NOVA IMS Information Management School

MGI

Mestrado em Gestão de Informação
Master Program in Information Management

Understanding musical genre preference evolution within a social network

João Pedro Ramos Pereira da Silva

Dissertation presented as partial requirement for obtaining the Master's degree in Information Management

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

UNDERSTANDING MUSICAL GENRE PREFERENCE EVOLUTION WITHIN A SOCIAL NETWORK

by

João Pedro Ramos Pereira da Silva

Dissertation presented as partial requirement for obtaining the Master's degree in Information Management with a specialization in Knowledge Management and Business Intelligence

Advisor: Filipa Alexandra de Peleja Madureira

December 2019

ACKNOWLEDGMENTS

In order to be able to write this thesis, several individuals helped me through a multitude of ways. As such, I believe it is only fair that, in this part of the work, I express how grateful I am for this help.

To my supervisor Professor Filipa Peleja, for helping me through this entire process, every step of the way, through multiple annotations, comments, corrections and ideas, I would like the express my gratitude.

A special thanks must also be given to Wen Dong and Todd Reid, for providing the social evolution data for both musical preferences and social relationships used throughout this thesis.

Last but not least, a very special thank you to my sister, for all the help during code creation, as well as to my parents, for the unassailable support throughout this entire process.

RESUMO

A música é um campo que simplesmente não pode ser desassociado dos aspetos sociais da vida. Durante a história da humanidade, a música mais popular consistiu sempre num reflexo dos diferentes aspetos da sociedade. Como tal, diferentes estudos foram feitos anteriormente que demonstram este reflexo e obtiveram diversas conclusões.

Nesta tese, iremos contribuir para este campo através de uma análise da evolução das preferências de géneros musicais ao longo do tempo através de uma rede social. Usando dados obtidos através de uma experiência de evolução social com cerca de 80 participantes faremos uma análise dos dados existentes. De seguida, esta análise é tida em conta para definir os princípios necessários para representar e analisar a rede social existente. Após esta definição, iremos avaliar a homogeneização da rede social ao longo do tempo. Isto é, iremos avaliar a evolução das diferenças de preferências musicais entre indivíduos que estão ligados na rede social, de forma a perceber se existe alguma tendência de estas diminuírem ao longo do tempo.

Um *Sequential Algorithm*, conhecido como *Hidden Markov Model*, é aplicado para prever mudanças nas preferências de géneros musicais, considerando as próprias preferências de cada individuo, bem como as preferências dos indivíduos com que este se encontra ligado na nossa rede social. O algoritmo *Support Vector Machines* é também utilizado para fazer o mesmo tipo de previsão que o modelo anterior servindo como comparação.

Por último, discutimos o processo e as limitações que conduziram à definição final do nosso modelo e de forma a contextualizar os resultados que foram obtidos através deste. Em suma, esta tese procurar acrescentar ao trabalho existente em termos de preferências de géneros musicais através de uma avaliação destes dentro do contexto de uma rede social e tendo também em conta a evolução destas ao longo do tempo.

PALAVRAS-CHAVE:

Preferências Musicais; Hidden Markov Model; Social Network Analysis; SVM

ABSTRACT

Music is a field that simply cannot be disassociated with the social aspects of life. Throughout human history, popular music has always been a reflection of the different aspects of society. As such, there is an interesting amount of studies available that showcase this reflection and draw multiple types of insights.

In this thesis, we will look to contribute to this field by assessing the evolution of musical genre preferences over time throughout a social network. Using data obtained through a social evolution experiment of around 80 different individuals we will make an initial assessment of our existing data. This evaluation is then taken into consideration in the next phase of our work where we define the principles necessary to represent and analyse the existing social network. Afterwards, we will showcase a representation of this network, as well as analyse it using various metrics and sub-structures commonly applied in Social Network Analysis. After this, we will evaluate the homogenisation of a network as time goes on. In other words, we will assess the evolution of differences in preferences between individuals that were connected in the social network, in order to understand if there is a trend of these differences diminishing over time.

A Sequential-Based algorithm, more specifically, a Hidden Markov Model is used to predict the change in musical genre preferences. This was done by considering each individual's own preferences as well as the preferences of his connections within the social network with the ultimate goal of assessing how influential the network is in the evolution of a person's musical genre preferences. To tackle the same research question and provide an alternative approach, as well as a comparison model, we used a Support Vector Machine model.

Finally, we discuss the results and limitations that led to our model definition. Overall, this thesis seeks to build upon previous work regarding musical genre preferences by assessing these within the context of a network and taking into account the evolution of these over time.

Keywords:

Musical Preferences; Hidden Markov Model; Social Network Analysis; SVM

INDEX

LIST OF FIGURES

LIST OF EQUATIONS

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

HMM Hidden Markov Model

SVM Support Vector Machine

SNA Social Network Analysis

1. INTRODUCTION

As time has gone on, the continuous progress in the Machine Learning (ML) field has brought on new insights on multiple fields, even in those that already have been studied for decades. This means that we can now look at reality from a standpoint where new ideas and progress can be made in areas that are usually considered unstructured.

An example of advances in ML is the impact in the music domain. For millions of people around the world, music is part of their daily life. From listening in their car, during their daily chores or even in the shower, it serves for many as a way to escape from the daily routine. However, the studies and papers relative to this particular art within social and personality psychology are few and far between. Of the nearly 11,000 articles published from 1965 to 2002 in social and personality journals, music was listed as an index term (or subject heading) in only seven articles (Rentfrow & Gosling, 2003).

Looking at the work published in the areas of social behaviour and music (Denora, 1999; North & Hargreaves, 2012; Saarikallio & Erkkilä, 2007), only a few take into account the correlation between social behaviour and musical preferences (Christenson & Peterson, 1988; Rentfrow & Gosling, 2003, 2006, 2007). While previous work has given us a solid perspective on this relationship we look to build upon this by taking into consideration the possibility that not only can your music preferences be connected to your social behaviour but also your social relationships. With that possibility in mind, various questions arise:

Do you tend to be friends with people with similar music preferences as yours?
Are there specific genres that are becoming more popular within your group of friends?
Does your taste in music change according to who are your closer friends?

These types of questions can be analysed in different ways for different environments. To best provide an answer, in the present work, our goal is to take advantage of the increasing amount of information available regarding social networks (Dong et al., 2012), through the use of statistical models. The objective is to test the hypothesis that music and the existing social connections have a relationship. For the outcome of our work we identify two different possible interpretations: the first is if both you and your group of friends will tend to have similar musical preferences as time goes on; while, the second is essentially that if a friend of yours starts having interest in different genres then you will also tend to have interest in those genres.

In this work we use survey data from 84 campus students obtained over a period of 2 years. The extend of this period allows us to perform a study where we can better understand musical tastes of each individual as well as model the social network itself. This data consists of students connections where for each member of the network the individuals he/she shares a relationship with are identified as well as what type of relationship it is. At the same time, it includes what genres of music each individual likes and the positiveness level for that preference, represented by a scale that describes slight, moderate and high interest.

The main objectives of this work are as follow:

- Creating and analysing the existing social network based on our data
- Assess how the existing musical preferences develop over time and if they tend to be more homogeneous within the social network as a whole
- Predict the evolution of musical preferences taking into account both the individual's preferences at a previous period of time, as well as the preferences of the individuals' social network

For the first objective, we will analyse the data from the social relationships survey in order to define the existing principles for the network, and then take advantage of Social Network Analysis (SNA) algorithms to assess all the characteristics of this network. SNA can be defined as a "*study of human relationships by means of graph theory*" (Tsvetovat & Kouznetsov, 2011). It is particularly suited for observing and studying patterns of sociality in order to better understand how all characteristics that define us impact these patterns, making it ideal for the goal of this work. As for the second objective we will take a deep dive within the survey data for musical preferences in order to detail the distribution of each genre preference within the network and conclude with an analysis over time of how each genre preference differences evolve within the network.

Finally, in terms of our third objective, we will employ a Hidden Markov Model (HMM) using variables that represent both the individual's preference, as well as the preferences of those that have relationships with him, and attempt to use these to predict genre preference evolution within the network. A Support Vector Machine model will also be used, using the same variables and for the same purpose, in order to provide a different approach and a suitable comparison.

2. LITERATURE REVIEW

In this work, while analysing the evolution of musical preferences throughout a network, we use various different concepts from multiple fields. As such, we wish to take a look at some of the existing literature regarding the key fields that serve as the basis of our work, whose concepts and conclusions we build upon. These fields consist of music and social behaviour, SNA and HMMs.

2.1. Music and social behaviour

In terms of the existing work linking music and social behaviour, there have been multiple studies that showcase existing links between these two. One of the examples for this type of work consists in exploring the role of music in adolescents by using data from group interviews and follow-up forms, analysed using constructive ground theory methods (Saarikallio & Erkkilä, 2007). Further work in terms of assessing the role of music in everyday life has also been done in a different paper that argues that the role of music changes as a result of the social and technological changes within music itself, discussing the different psychological functions it has in cognitive, emotional and social domains (North & Hargreaves, 2012). Another study also uses data from interviews in order to showcase how the interviewees use music as resource for the conduct of emotion, increasing or lowering their energy levels (Denora, 1999).

Focusing on existing work regarding social behaviour and music preferences, there have been a few studies that have made this type of connection. One of these studies consists in analysing the existing music preferences within various college students, based on responses regarding music use and preference levels, concluding that the existing underlying structure of music preferences has to take into account various factors and that there are relevant differences between males and females in terms of these structures (Christenson & Peterson, 1988). A different work focused on analysing the correlation between music preferences and personality was also done. Using data from 6 different studies, various different music-preference dimensions were obtained in this work, showcasing that the correlation mentioned does exist to a certain level (Rentfrow & Gosling, 2003). Another paper tackles this topic from a different perspective by showing that different individuals use their various musical preferences to relate information regarding their personalities to others, and that the recipients of this information can use this to form impressions of other individuals. This was done by demonstrating, through different studies, that music is an important topic of conversation between strangers, and that observers could form consensual and accurate representations of an individual based on its musical preferences (Rentfrow & Gosling, 2006). A year later, the same authors pose a similar question but from a different perspective, focusing on the validity of existing stereotypes of fans of different musical genres, namely in terms of personalities, personal qualities, values and alcohol/drug preferences. Taking into account a similar dataset from the previous paper, the authors found that individuals have stereotypes regarding each musical genre, that are strongly solidified, and that these have a basis of truth (Rentfrow & Gosling, 2007).

With these studies in mind, we concluded that there is a solid basis in terms of existing work regarding these concepts, since there have been multiple studies that showcase links between musical preferences and social characteristics. We will look to add to these studies by taking into account the change of musical preferences over time, as well as assessing the influence of an individual's relationships in a social network upon these.

2.2. Social Network Analysis (SNA)

The definition of SNA is the study of social relationships (groups) by representing these as networks of individuals connected by social relations. A network consists of a modulation of a system comprised of specific individual nodes and their ties (connections). When analysing these networks, the term "node" can be replaced with the term "individual", while the term "tie" can also be replaced by the word "relationship". A "Group" is defined as the complete network, or rather the sum of all individuals that can potentially interact and that can be differenced from other such sums of individuals. The smallest amount of ties between two nodes is essentially the length of the path from one node to another (Wey, Blumstein, Shen, & Jordán, 2008; Ferreira, 2013).

2.2.1. Birth and basics

Social Network Analysis (SNA) has been used for a variety of different applications, and much of the previous work that uses this has been focused on obtaining a better understanding of the multiple existing behaviours and social relationships by analysing the relationships between different individuals (Ferreira, 2013)
Measuring social relations started with questions such as:

Q1 How much time does person A spend with person B?

Q2 What about the time that person A spent with others, such as person C and D?

Q3 If, for person A, the time spent with C is bigger than the time spent with D, what is the significance of this and how can we measure it?

Q4 Considering these relations, how can you define the "spatial distance" between these 4 individuals during similar situations and what role does this distance have upon different behaviours and decisions?

These are among some of the issues tackled by Moreno in 1953 (Freeman, 2004; Ferreira, 2013). Since this, different authors, such as Harrison C. White, researched and developed theories about groups of individuals that worked together (Freeman, 2004; Ferreira, 2013), such as modelling stochastic processes and giving a more adequate account of complex social structures. These early papers are what can be considered as the birth of SNA.

2.2.2. Other applications – Using SNA to explain different types of behaviours

The nature of these types of behaviours can vary significantly. One study, for example, focuses on the spread of infection through a social network. In this study a model is used to predict the spread of infection on an individual level and gives useful information in terms of

existing infection paths (Dong et al., 2012). Another study uses data to show the evolution of individual behaviour and co-relations in a student dormitory. This behaviour varies from discussing politics to interacting on Facebook and Twitter. By modelling the existing social dynamics, they demonstrate the ability to predict friendship, and can synthesize useful and accurate behaviour and interaction projections (Dong et al., 2011).

The shaping of eating behaviours (and, as a consequence, body weight) has also been studied using this discipline, particularly by synthetizing evidence of associations between this kind of behaviours and young people's social networks. It was demonstrated that school friendships are critical in shaping young people's eating behaviours and bodyweight and/or vice versa, and suggests the potential of social-network based health promotion interventions in schools (Fletcher, Bonell, & Sorhaindo, 2011).

An interesting demonstration was also done in a paper that studied human mobility patterns combining the temporal information about the whereabouts of users with information on the types of places they visit. Interesting associations (e.g. very strong weekly patterns, increase of activity as the week progresses) were discovered, that allowed to cluster individuals based on their behaviour (Preoţiuc-Pietro & Cohn, 2013)

2.2.3. Relevant Concepts

Some of the most important information regarding a social network can be obtained when analysing some of the most commonly used SNA metrics that exist, such as degree, eigenvector, closeness, and betweenness centrality (Arif, 2015).

A centrality measure essentially assesses information regarding how important nodes and edges are within the network. Starting by looking at degree centrality, it is considered the "simplest of all the centrality measures" (Arif, 2015). For a given node v in an undirected network, such as the one we will use, it represents the number of links it possesses and is usually used to identify the nodes that possess the highest number of connections within the network. A degree $C_d(v)$ for node v can be calculated using Equation 1.

$$C_d(v) = \deg(v)$$

Equation 1 - Degree calculation for node v

Where $\deg(v)$ is the number of edges connected to node v

The closeness centrality measures the "degree of nearness" between a specific node and the rest of the nodes within the network (Arif, 2015). It is calculated through inversing the sum of the smallest distance between a specific node and the rest of the network. Thus, the higher the closeness centrality, the more central this node is within the network. For a given graph G with n nodes, the closeness centrality of a node v can be measured using Equation 2.

$$C_c(v) = \frac{n-1}{\sum_{k=i}^{n} d(u_i v)}$$

Equation 2 - Closeness calculation for node v

where $d(u_i v)$ consists of the geodesic distance between u_i and v

The betweenness centrality represents the "fraction of all shortest paths that pass through a given node" (Arif, 2015). In short, it measures how many times a node behaves as "bridge" in the shortest path between two different nodes. A node that possesses high betweenness centrality tends to play crucial roles in the network, since they have an important role in the "cohesiveness" of the network, they are considered central and indispensable to it. This centrality for node v can be calculated using Equation 3.

$$C_B(v) = \sum_{s \neq v \neq t \epsilon V} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

Equation 3 - Betweenness calculation for node v

where $\sigma_{st}(v)$ is the total of shortest paths from node s to t and $\sigma_{st}(v)$ is the number of paths that pass through v.

Finally, the eigenvector centrality is considered a more "sophisticated version of degree centrality" (Arif, 2015), since it takes into account the number of links a node is part of, but also takes into account the quality of these. In this measure, having connections with nodes that possess a bigger amount of connections leads to a higher centrality value. Variations of this eigenvector centrality include Google's PageRank and Katz Centrality (Arif, 2015).

Let's assume $A = (a_{v,u})$ as the adjacency matrix of a graph G with V nodes and E connections. Then A can be defined as:

$$A_{v,u} = \begin{cases} a_{v,u} = 1. if\ node\ 'v' is\ linked\ to\ node\ 'u' \\ a_{v,u} = 0, otherwise \end{cases}$$

The eigenvector centrality of node v can be defined using Equation 4.

$$C_E(v) = \frac{1}{\lambda} \sum_{u \epsilon N(v)} x_u = \frac{1}{\lambda} \sum_{u \epsilon G} a_{v,u} x_u$$

Equation 4 - Eigenvector centrality for node v

Where $N(v)$ represents the set of neighbours of node v (other nodes that node v is connected to) and λ is a constant

Another way of analysing existing relationships within the network is by assessing various different structures within the network. One of these types of structures consists of a subgraph. A subgraph is a subset of the nodes of a network, and all of the edges linking these nodes. Any group of nodes can form a subgraph, meaning there are several interesting ways to use these (Tsvetovat & Kouznetsov, 2011)

Component subgraphs (or simply components) are parts of the network that are not connected with each other (Tsvetovat & Kouznetsov, 2011). For example, before the meeting of Romeo and Juliet, the two families were not connected in any way (save for the conflict ties), and thus could be assessed as components.

One method for creating these components is the Island Method (Figure 1). It has this name since it uses a defined metric, such as edge weight, as "water level", obtaining "islands" (components) based on a defined threshold. To implement the island method we need a function to virtually raise the water level. The function takes a graph, and applies a threshold ("water level"), letting all edges above a certain value through, and removing all others. Multiple iterations can be made with different thresholds in order to identify the optimal structures for the problem analysed. It may take some trial and error, but a well-tuned "water level" can provide a very insightful analysis of a large network—instantly obtaining the cores of the most activity (Tsvetovat & Kouznetsov, 2011).

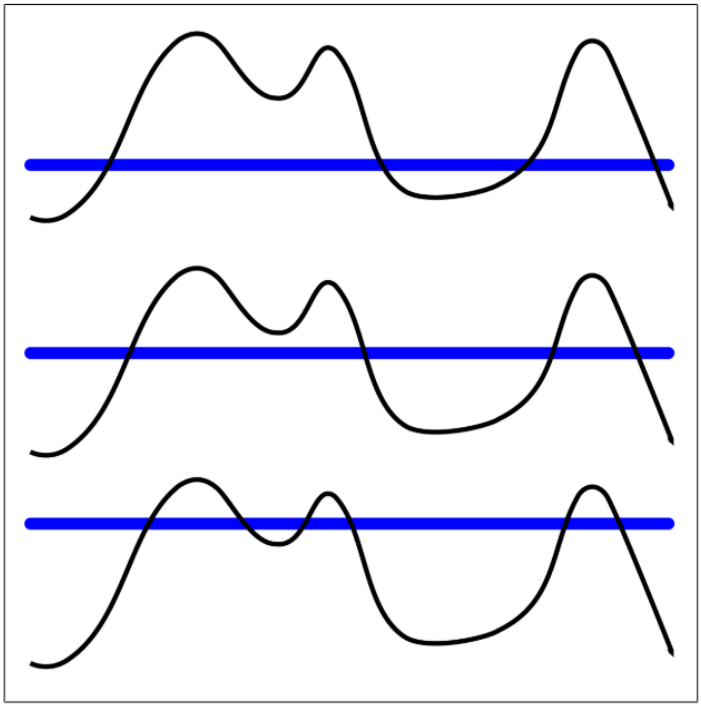

Figure 1 - Applying the Island method using different water levels (Tsvetovat & Kouznetsov, 2011)

Another method is by identifying and isolating Ego Networks. To do this you must apply breadth first traversal, a method for defining a social network, where you start with a single node, visit all of the immediate neighbours, and only afterwards visits their neighbours'

neighbours. The main difference being that the defined radius of the search is small, usually not more than 3 (Tsvetovat & Kouznetsov, 2011). These could, in our opinion, be very helpful if you apply this technique using as the initial node members of the network the ones that are highly influential.

Another interesting structure are triads (Figure 2). These consist of simply three nodes connected in some way. However the analysis of these can vary a lot, due to the various different types of triads. Taking into account all different possible connections between 3 people, there are 16 types of possible directed triads.

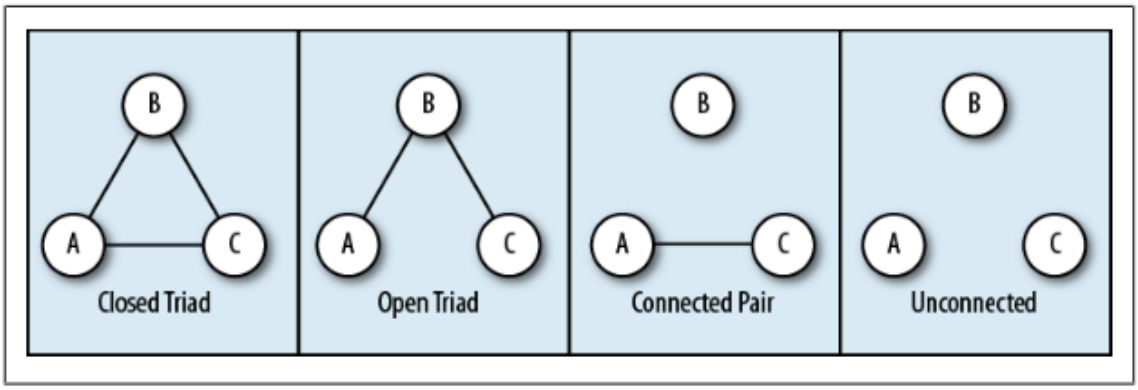

Figure 2 - Different types of triads (Tsvetovat & Kouznetsov, 2011)

An interesting study using triads was done by Newcomb (Tsvetovat & Kouznetsov, 2011). In this paper, 17 students, all of which consisted of white men, were recruited so that they could live in a frat house for a semester, in exchange for providing their personal data. Every week, researchers had interviews with every single one of the students asking them to attribute to their various interactions with the different fraternity brothers a rank, ranging from 1 (best) to 16 (worst) (Tsvetovat & Kouznetsov, 2011). The findings of this study were that triadic structures provided the most stability over time, with students resolving conflicts, organizing different types of events together, and ultimately having the most control in the interactions among the various fraternity brothers. With this in mind, this was a particularly interesting study, since it analyses essentially the same demographic that we analyse in this thesis.

Cliques are another interesting structure type. A clique is defined as a "maximal complete subgraph of a given graph"—which means these consist of groups of individuals where everybody is connected to every other individual. By using the word "maximal", we imply that if we were to add any new node to a clique, the structure would no longer be completely connected. This means that, in essence, a clique is a combination of several closed triads that overlap, inheriting many of the "culture-generating, and amplification properties of closed triads" (Tsvetovat & Kouznetsov, 2011).

The next class of structure we wish to analyse are clusters (Figure 3). The number of possible clustering algorithms is very large and varies wildly. A key concept for the structures are the functions that allows you to measure "distance" between points. The definition of this distance can vary significantly, from geographical to time based distance. Within the scope of social networks, one of the most useful distance definitions that can be used is the path length

between nodes (Tsvetovat & Kouznetsov, 2011). An example of a clustering algorithm consists of hierarchical clustering, in which, based on the distance defined, different nodes from the network can be grouped into different clusters at different levels. A clustering threshold is then defined, through which the final clusters are chosen depending on what is most useful.

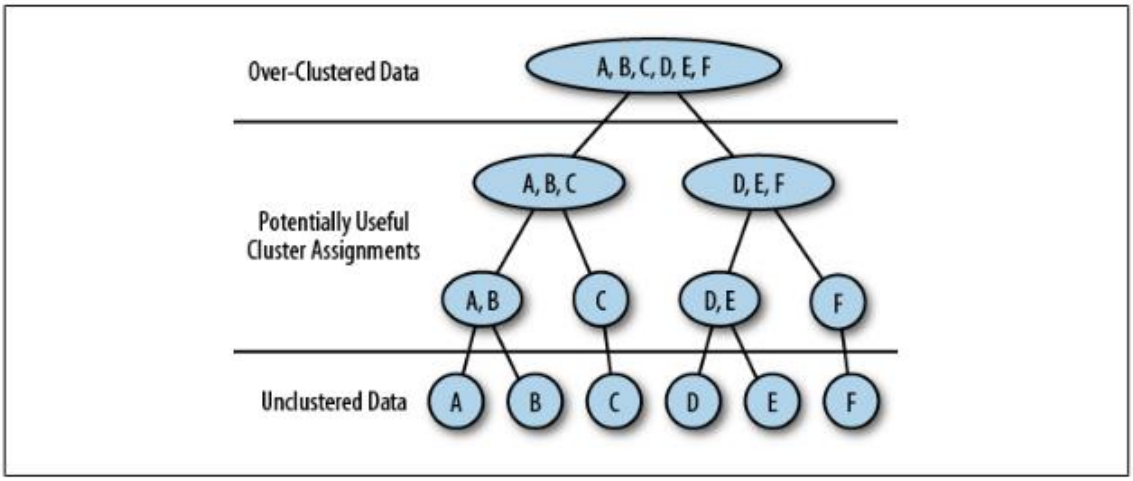

Figure 3 - Examples of clustering (Tsvetovat & Kouznetsov, 2011)

The final example of a structure that we wish to mention is what is known as a block model. A block model is a simpler version of a network that is derived from a previous original network, where the existing nodes from a cluster are considered a singular node, and the existing relations between the original nodes are aggregated, becoming relationships between blocks (Tsvetovat & Kouznetsov, 2011). Like ego networks, block models are, in our estimation, particularly interesting if you apply it using the most relevant members of the social network.

2.3. Hidden Markov Model

2.3.1. A practical example

This model will be further explained later on during our methodology, but, for purposes of contextualization, a basic explanation can be given through a simple example. Imagine that, 30 years from now, due to lack of records and memory, you wish to recall what type of weather (cold, mild, or hot) there was at a particular time. While you have no specific information regarding the type of weather there is, you do possess specific records entailing which type of clothes you wore in a specific day. In order to decode the existing weather at each particular time, a simple approach would be to create a conditional model that, for example, if you wore a specific type of coat with a particular type of scarf, it indicated that the weather that occurred was cold. However, as you might guess, a process such as the one described above is a somewhat random and very ineffective way to solve this problem. Fortunately, however, these kind of problems are where a HMM excels (Da Silva, 2014).

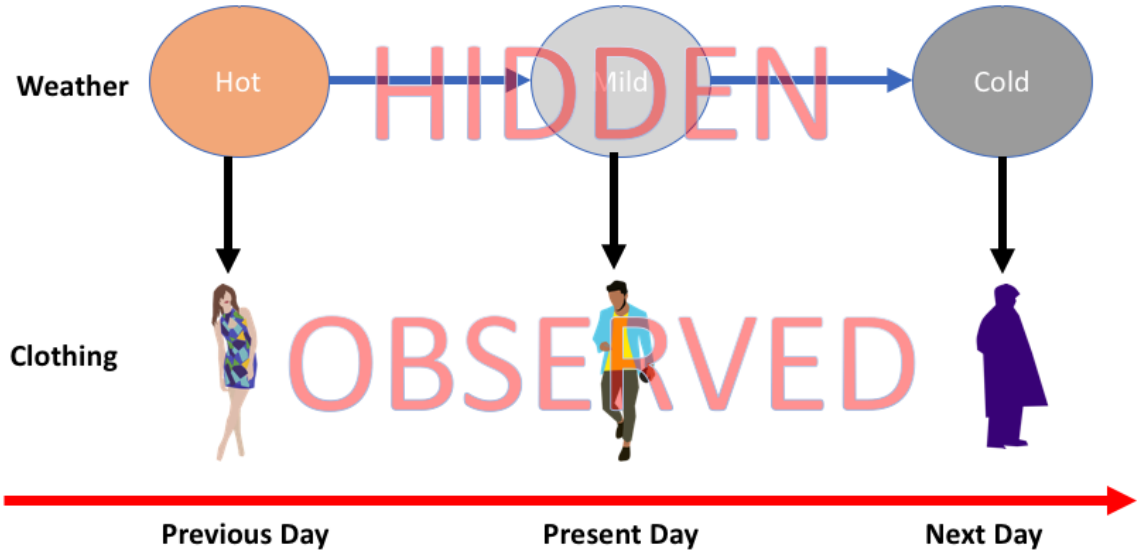

Figure 4 - HMM example ("HMMs Simplified - Sanjay Doairaj - Medium," n.d.)

Let's consider three hidden states, which will consist of the weather being hot, mild and cold. By assigning a probability matrix to both the possible combinations of clothes within my wardrobe at the time, as well as the three states, HMMs are capable of outputting the optimal state sequence by maximizing the probability of a sequence being produced by the model (Da Silva, 2014). Despite this model being unable to tell you which particular state occurs at the same time as each outcome, it is possible to infer these by looking at the characteristics of each state. It is logical to assume, for example, that specific clothing items would be used for cold weather, others for mild weather, and other different ones for hot weather.

2.3.2. Common Applications

In terms of applications, one of the possible uses for HMMs consists in part-of speech tagging (Kupiec, 1992) . In this work, a robust system designed for the aforementioned speech tagging was created based on an HMM, ultimately obtaining a model that correctly tagged 96% of the text tested. This model has also been taken into account for text segmentation and event tracking (Yamron, Carp, Gillick, Lowe, & Van Mulbregt, 1998). Within this paper, "stories" that relate to specific events were obtained, using a methodology based on HMMs. Another interesting type of application of this type of model occurs when using financial data. This model has been used for example, in attempting to assess how to successfully describe stylized facts of prices returns (Rydén, Teräsvirta, & Åsbrink, 1998) as well as predict the sign of financial local trends (Bicego, Grosso, & Otranto, 2008).

Despite not finding a specific work that uses this type of model to assess musical preferences throughout a network, there has been precedent in using a version of a HMM in order to infer infection on an individual level, based on the various flu symptoms throughout the members of the network (Dong et al., 2012). In this work, also mentioned previously as an example of use of Social Network Analysis, graph coupled HMMs are applied to assess spread of infection throughout a social network.

3. DATASET ANALYSIS

In order to be able to accomplish our goals in this work, we need to take a deep dive into the existing data that will be used. With this in mind, we start with a description of its characteristics, such as its original goals and the description of the individuals present within it. This analysis is key to our work, since it will be the basis upon which we will train our models and provide the characterization of our social network, as well as assess if there is any indication of whether there is a tendency for homogenization of musical preferences through the network

3.1. Dataset Origin

The original dataset from Wen Dong and Todd Reid consists of a Social Evolution experiment designed to closely assess the day-to-day life of an entire undergraduate dormitory, so that social scientists can assess their own models against the different spatiotemporal patterns and behaviour-network co-evolution contained in this data. This experiment covered the various phone calls, locations and proximities of the majority of residents who lived in the dormitory (over 80 %) used in the Social Evolution experiment, as captured by their cell phones from 2008 to 2009. The dormitory has a population of about 30 freshmen, 20 sophomores, 10 juniors, 10 seniors and 10 graduate student tutors.

This experiment was designed to capture the adoption of several different types of data, particularly, political issues, interpersonal relationships, epidemiological contagion, depression and stress, diet, exercise, obesity, eating habits, political opinions and privacy.

The collection of this data includes location, proximity, and call log, obtained through a cell-phone application that scans nearby Wi-Fi access points and Bluetooth devices every six minutes — through which the latitudes and longitudes of the Wi-Fi access points are referenced. Survey data includes political opinions (democratic vs. republican), smoking behaviour, attitudes towards exercise and fitness and diet, actions regarding academic performance, confidence and anxiety levels and, more importantly for our analysis, a sociometric survey for relationship analysis and a music preference survey using a wide assortment of genres.

Our work focuses on two specific features from the described dataset. The first one describes the interpersonal relationships and the second one consists of the existing musical preferences for each person. In terms of our data regarding relationships, these consist of either connections through Twitter or Facebook, being someone who discusses politics, someone who the individual socializes twice per week, or being a close friend. In terms of musical preferences, these consist in preferences regarding 11 genres: Classic Rock, Classical, Country/Folk, Heavy Metal / Hardcore, Hip-Hop / R&B, Indie / Alternative Rock, Jazz, Pop / Top 40, Showtunes, Techno / Lounge / Electronic and Other (representing music that doesn't fit any of the previous genres). These preferences can vary from 3 different levels: slight interest (1), moderate interest (2) and high interest (3).

3.2. Musical Preferences Dataset

Looking at the survey questions related to musical preferences we have 248 responses from 79 individuals with a total of 1,527 entries. This amount of entries is due to the fact that, for a given response, each individual can demonstrate interest in multiple genres. In terms of the number of responses per month, from Table 1 we can observe that, while in most months the number of responses is relatively stable (between 20 to 30% of total responses overall), there are three specific months (November, December and March) which seem to present themselves as non-representative. These months do not represent our population well due to the small amount of responses given (6, 4 and 2 responses, respectively). However, from a yearly perspective, the number of responses is relatively balanced, with only a slight skew towards the responses in 2008 (56.45%) rather than 2009 (43.54%)

Unique Responses	Sep 2008	Oct 2008	Nov 2008	Dec 2008	Mar 2009	May 2009	Jun 2009	Total
Number	70	60	6	4	2	54	52	248
Percentage of total	28,22%	24.19%	2.41%	1.61%	0.81%	21.77%	20.97%	100%

Table 1 - Unique responses of musical preferences survey per month

In order to have a better understanding of our data, a deeper look must be given to the months that have the lesser amount of responses. In November, every answer given was in fact the second response to the survey of that particular individual. This contrasts with data from December 2008 and March 2009, where every single answer was the first for that specific individual. This analysis might seem trivial, but will prove crucial during our data preparation, as we define what we consider the definition of a state for our Hidden Markov Model.

Another important analysis consists in how many genres does an individual usually have interest in at a particular time. This is very much relevant due to the fact that, since we will be attempting to analyse change of preferences over time, we will need to assess what we consider an individual “changing preferences”. However, this is only possible if we assess the genre frequency per response of a survey. Overall, we found that an individual possesses, on average, interest in about 6 different genres. In most months where this survey was presented, the number of genres each person has an interest in is closer to 7/8 different genres. The big exception is the month of May, whose number of genres that an individual has interest is of around only 2 different genres. Taking a closer look at data for this month, it seems that each individual only responded a max of three different genres with interest. Since we do not have access to the surveys themselves or the individuals that took them, we can only speculate, but one of the possible explanations for this abrupt change could be a structural difference in the survey given in this particular month, that induced the individuals that took the survey at this particular time to only put at most three different genres in their responses

In terms of how many times each individual responded to this survey, on average, there were around 3 responses per individual. In terms of response timing, in September, as

the first month in which the survey was responded to, only possesses answers which consist of each individual's first response, while in October the answers are either the first or the second response. Also, as expected, in May, the responses given are either the second and third responses of an individual, while, in June, these can be either the third or fourth responses to the survey.

In terms of the genre preferences themselves, in order to provide a more interesting analysis, we focus on looking at these from a yearly perspective, since a monthly perspective would be heavily skewed in the existing months with particularly low amount of responses.

As you can see from Table 2, despite being slightly skewed to indie/ alternative rock, classic rock and classical, the three most popular genres, others are also relatively well represented, since none of them drop below 5% of representation of our dataset at a particular year and are relatively even in terms of the genre preference.

Genre	2008 (%)	2009 (%)	Total (%)
Classical	11,49	11,22	11,39
Classic Rock	10,99	12,38	11,46
Country / Folk	7,03	6,38	6,81
Heavy metal / Hardcore	7,52	7,16	7,40
Hip-hop / R&B	7,52	5,61	6,87
Indie / Alternative Rock	10,89	14,70	12,18
Jazz	9,70	9,09	9,50
Other	7,13	6,00	6,74
Pop / Top 40	9,40	10,25	9,69
Showtunes	8,51	7,54	8,19
Techno / Lounge / Electronic	9,80	9,67	9,76

Table 2 - Relative frequency per timeframe and genre

Another important conclusion we can obtain from Table 2 is the fact that there are no significant shifts in genre interest throughout the time of this study. This is shown through the fact that the Top 3 most popular genres (indie/alternative rock, classic rock and classical) and the 3 less popular genres (country/folk, hip-hop/r&b and other) remain the same in 2008 as well as 2009 – despite the order of most/less popularity between them changing. However, there is some variation of the musical preferences, both when taking account the absolute and relative frequency. Examples of this are the variation percentage wise in the Indie/Alternative Rock genre which goes from representing slightly above 10% of all genre preferences in 2008 to around 15% in 2009, and in the Hip-hop/R&B genre that shifts from having around 7.5% representation in 2008 to around 5.6% in 2009. These variations lead us to the conclusion that there are in fact changes in preferences along the network, which begs the question: do these changes have any relation to the existing social relationships at the time?

Looking now at the correlation between genres, we wish to possess a deeper understanding of our musical preference data, by assessing whether or not there are any specific genres that have any significant amount of correlation. This is very much relevant since during our experimentation with our HMM, which is later discussed, one of the

possibilities we pondered was doing some type of feature reduction. If we possess multiple musical genres that have strong correlation, one possible hypothesis could have been using only one of these during our HMM data preparation.

This correlation was measured by using the Pearson coefficient. This coefficient is one of the most commonly used correlation coefficients and consists of a measure of the linear correlation between two different variables (Gingrich, 2004). This is calculated by dividing the covariance of the two variables analysed (in this case, the joint variability of two genres), and dividing by the product of the standard deviation of each variable. If these variables have a linear relation in the positive direction between each other, then this coefficient will be positive and considerably above 0. However, if this relationship is in the negative direction, so that when one variable occurs, the other does not, then this coefficient's value is negative. The values of Pearson' coefficient can range from -1 to +1, with values approximating 0 signifying a lack of a relationship between the two variables.

In terms of assessing the results, the dataset is fairly balanced since there are no very high values between genres (the highest is of around 54 % percent, while the lower ones are of around 10%). The Classical genre is particularly unique since, other than its correlation with Jazz, it possesses the lowest correlation coefficients, which signals a more separate community that enjoys these genres more exclusively.

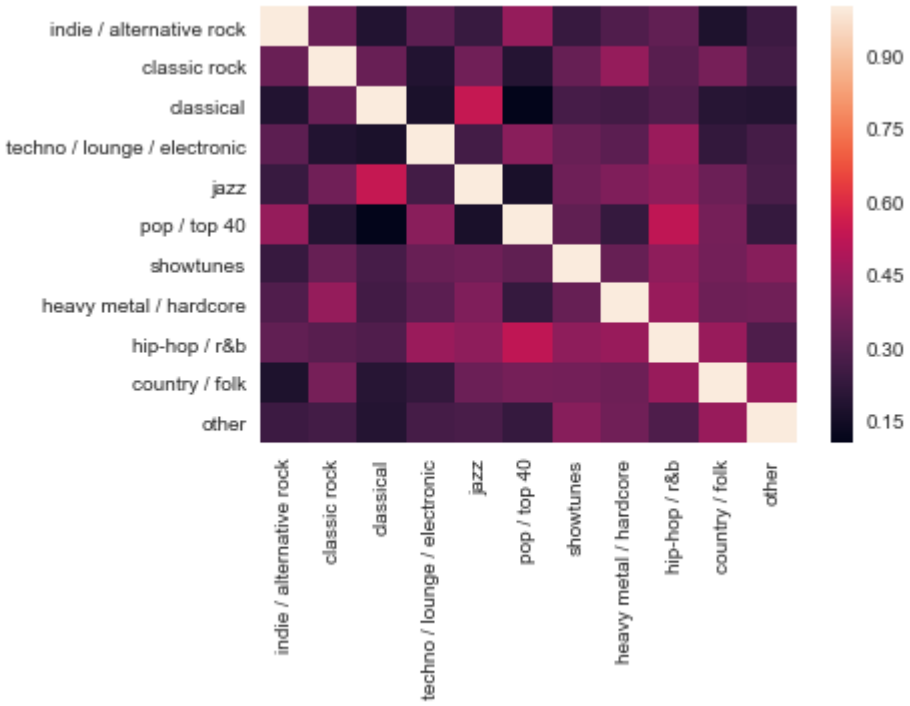

Figure 5 - Correlation matrix between different music genres

3.3. Social Relationships Dataset

Now that we have a better understanding of the existing dynamics within our musical preferences' data, the next step is to analyse the behaviour of our existing data regarding social relationships. This dataset possesses 392 responses, corresponding to 80 individuals, totalling for 31,918 rows. This significant number of rows is due to two factors. The first one, is that, in this survey, each individual can represent various other individuals with whom that person possesses relationships with. This is logical, since it represents the number of connections that an individual possesses within the network. The other reason, however, is that an individual can represent multiple relationships with another particular individual. For example, individual A can represent in a particular survey that individual B is a Facebook friend, that they socialize twice per week and that individual B is a close friend. This means that during our network definition, we will not be able to simply consider a mention of an individual in a survey response as a connection, since this would lead to multiple duplicate connections.

In terms of an assessment of responses over time, and unlike the musical preferences dataset, this dataset is effectively balanced on a monthly basis, since there are no months that possess a significant lesser amount of responses. This is shown in Table 3, which showcases that the number of responses per month only varies between 15 to around 18% of the total number of responses. From a yearly perspective, however, the dataset is very much similar to the one regarding musical preferences where the number of responses is slightly higher in 2008 (52.81%), rather than 2009 (47.19%)

Unique Responses	Sep 2008	Oct 2008	Dec 2008	Mar 2009	May 2009	Jun 2009	Total
Number	71	67	69	62	62	61	392
Percentage of total (%)	18.11	17.09	17.60	15.82	15.82	15.56	100

Table 3 - Unique responses for social relationship survey per month

Another important analysis, similar to the one given in the musical preferences dataset analysis, is the number of existing responses to the survey per individual. In this dataset, each individual on average responded 4.9 times. This seems logical however, when considering data from Table 3, since we have 80 individuals that responded to this survey and there is no month that had less than 60 unique responses. It also means that, each month, at least 75% of the total network gave a survey response. In fact, the reason why this average isn't closer to being 6 responses, is the fact that there are 13 individuals that responded in 2008 but not in 2009, and 2 individuals that responded in 2009 but not in 2008. Since these did not respond more than 3 times, it heavily skewed the average mentioned previously. Taking this previous analysis in consideration, an issue arises: How can we define what is the network of an individual at the time in which he took a musical preferences survey when the timing of these surveys, and the surveys related to an individuals' social relationships, do not align? There are multiple individuals that responded to musical preferences in November, but there are no responses for the social relationships survey during this month.

Relationship	2008	2009	Total
FacebookAllTaggedPhotos	29,95%	31,60%	30,77%
BlogLivejournalTwitter	28,76%	30,07%	29,41%
SocializeTwicePerWeek	19,44%	16,25%	17,85%
PoliticalDiscussant	13,90%	13,87%	13,89%
CloseFriend	7,95%	8,20%	8,07%

Table 4 - Relative frequency of the types of relationships represented in social relationships survey

Focusing now on the type of relationships documented within the survey, we can see that relationships that were documented seem to be mostly through Facebook or Twitter (around 30 % of overall representation for each), followed by the type of connections that represent socializing twice per week and political discussant (between 14 to 20 % each, depending on the timeframe analysed). The rarer relationship is, as expected, being close friend, which consists of around 8% of all relationships.

In terms of differences of types of relationships over time, the dataset seems to be relatively stable since differences between 2008 and 2009 in the number of relationships documented per relationship type only tends to vary around 1 to 2%, with the exception of the relationship that represents socializing twice per week whose relative frequency is of around 3% less in 2009 than in 2008.

4. NETWORK DEFINITION, REPRESENTATION AND ANALYSIS

Now that we have an understanding of the particulars regarding our social relationships data, we have the necessary information to define the principles of our network based on this data. After this definition, we will showcase the representation of this network, analyse it by computing the various SNA metrics, as well as assess the possible structures within a network mentioned in our Literature Review.

In order to define our network, we must take into consideration some of the issues mentioned during our analysis of the dataset. First off, one of the particulars of our data is that, within a specific survey taken by an individual, this person can represent multiple different relationships types for the same individual. This means that, for example, if we use an input of individual A for individual B as a connection, this could lead to duplicate connections between individual A and B. In order to solve this, we ensure that, for a connection between two different individuals, we always take into account the relationship type that seems to reflect the "strongest" type of relationship. In our estimation, this makes logical sense since, for example, a relationship between two persons that are close friends should be reflected as such, instead of merely being connections through Facebook. This definition implies a different question. What do we consider as different levels of relationship "strength"? Considering the distribution mentioned in Table 4, as well as what we believe are the logical associations, we considered being close friends as the strongest type of relationship, either being a political discussant or socializing twice per week as the middle level of relationship type and connections through Facebook or Twitter as the weakest relationship type. While not explicit in our network analysis, these characterization levels will later be taken into account during the experimentation part of our work for our Hidden Markov Model.

Another issue that we found during our data analysis was the fact that even though each individual responds to the musical preferences survey on average 3 times, they respond to the social relationships' survey nearly five times. There were also individuals that responded to one of the surveys in a particular month in which no data for the other survey existed. This means that we can't fully allocate each specific snapshot we possess of the existing social networks to a specific snapshot we possess of the existing musical preferences. With this in mind it was chosen to handle our network as static over time. Therefore it is not taken into account the evolution of the network over time. Instead, we take all unique connections between individuals through all surveys and, as mentioned, select the strongest level of connection. While we understand that taking into account this evolution of the network in a different manner could have provided additional benefits (such as assessing whether the network changes according to the existing musical preferences and better prediction capability within our HMM, which will take into account this network) when facing the problem mentioned, and considering the stability of the network, as seen through our previous analysis in Table 3 as well as Table 4, we found this approach as a fair compromise that would not heavily impact our assessments going forward.

Our final decision regarding our network consisted in deciding whether our network would be undirected or directed. A directed network is essentially a composition of nodes linked through directed edges, where each node is a link from one node to another, with each specific direction being important (Easley, 2010). This contrasts with an undirected network, in which there is no specification regarding direction of any links within a network. Since our data essentially consists of surveys where an individual represents how another individual relates to him, whether be a mere Facebook connection or a close friend, our first idea consisted in assessing our network as directed. However, when taking a closer look at some of the SNA metrics of this network, we found that, for specific individuals, either the in-degree or the out-degree was 0. As explained previously in Chapter Literature Review, the degree centrality of a node is essentially the number of connections of that node within a network. However, if we are considering a directed network, there are two different types of degree, the aforementioned in-degree and out-degree. The in-degree consists of the number of incoming neighbours (Butts, 2008), meaning it is a type of degree that only takes into account connections that are directed towards the particular node being analysed. The reverse is true for the out-degree, it only considers connections of a node that are directed to different nodes other that the node being analysed. Therefore, for these specific individuals, they either are mentioned but do not mention anyone in the existing surveys, or are not mentioned by anyone that participated in the surveys but do mention others. Since we do not possess the details of the survey carried out, we can only speculate the reason for this, but one possible reason could be due to lack of survey participation or survey limitations (individuals that were not mentioned by others might not have been included in the list of individuals a person could mention within the survey). As such, keeping our network as directed would lead to loss of data when applying it in our HMM and SVM. Data regarding individuals that did not respond would be left out, while data for individuals that were not mentioned by others would be forgotten in other individuals' connections. Having this in mind, and despite conceding that an analysis of this network as directed could provide interesting results, in order to avoid this loss of data (which is already very limited, as explained in Chapter Experimentation and Discussion), we will assess our network as undirected.

Now that we have taken all the necessary steps towards the definition of our network, we can take a look at its representation, present in Figure 6.

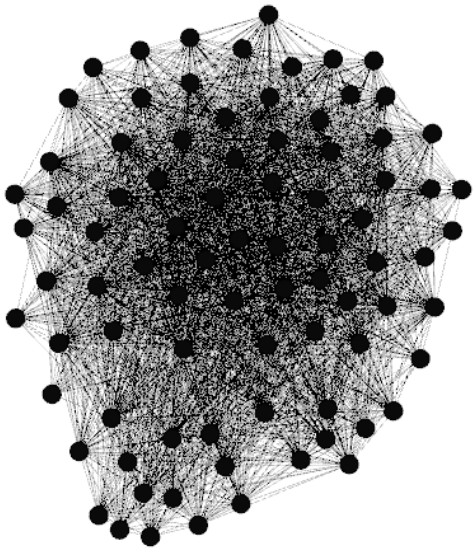

Figure 6 - Representation of existing social network

Looking at this network, it is easily observable how connected it seems to be, since every node exhibits multiple connections and the central nodes being particularly connected with a very large amount of other nodes. Another interesting representation is assessing the network together with our musical preferences' data (Figure 7).

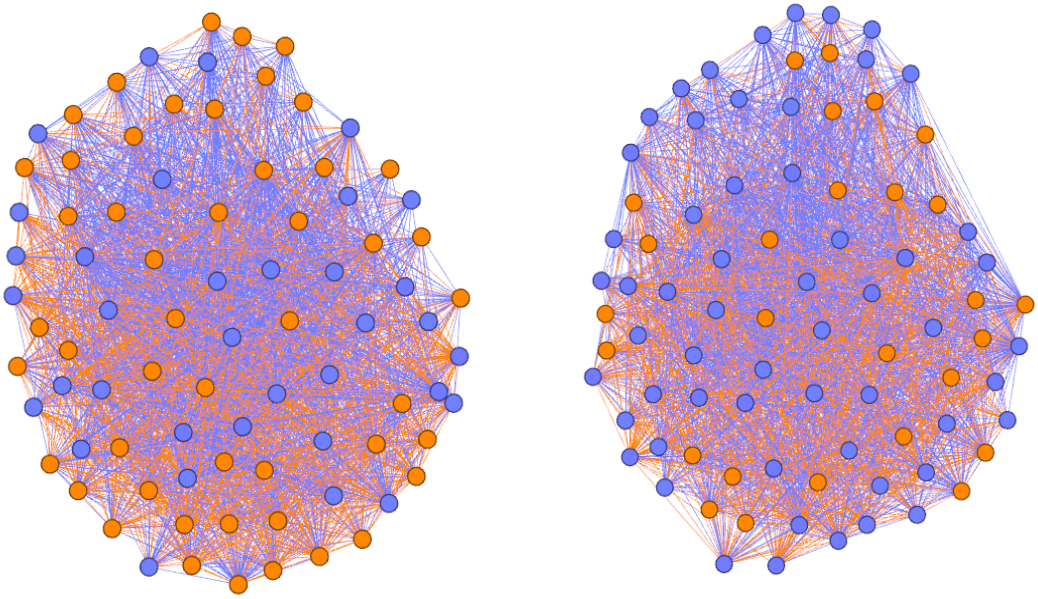

Figure 7 - Representation of our social network, coloured according to genre interest for hip hop / r&b in 2008 (on the left) and 2009 (on the right)

In Figure 7, you can see an example of musical interest changing over time, within the context of a social network. You can find the changes of genre in hip hop / r&b from 2008 to 2009, in which blue is not preferred and orange is preferred. A particular change occurs, for

example, in individuals 33 and 34, present in Figure 8, that had interest in the hip-hop / r&b genre in 2008 but no longer had this interest in 2009.

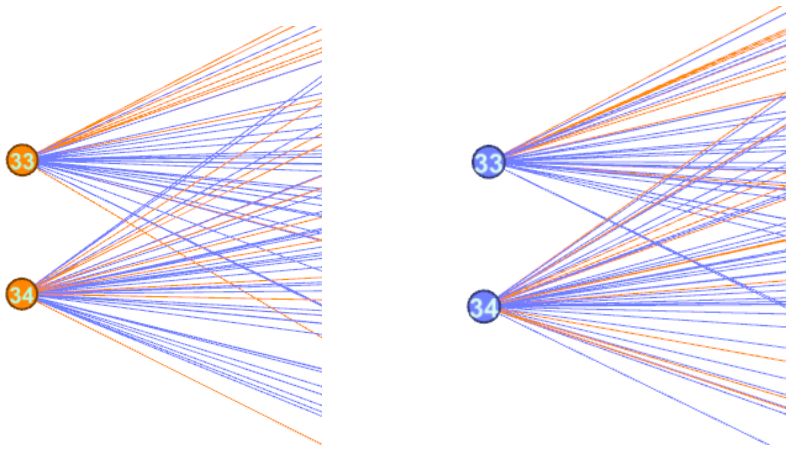

Figure 8 - Representation of genre interest variation of ids 33 and 34 within our network

The next step will consist in analysing this network. To do this, SNA centrality metrics such as Closeness, mentioned in Chapter Literature Review, will be computed to support our analysis. It is important to mention that all metrics used follow the previously mentioned equations, with the exception of degree centrality, that is normalized by dividing the absolute degree by $n-1$, in which n represents the nodes of the network. This was done to provide a better assessment of the network by taking into account the degree values in a relative, rather than absolute, fashion.

When taking into account the existing results in Figure 9, in terms of degree centrality, we can conclude that the previous intuition done through assessing our social network representation was correct. Every single individual is strongly connected to our network as a whole, having connections to at least 26 % of the network. Moreover, there is a significant number of individuals that have connections to all or nearly all of the individuals of the network. This incredibly high amount of connections ultimately influences all the other metrics. Considering closeness centrality, its values are also extremely high, with many ids having a closeness of 100% and no individuals with a closeness lower that 57%. This means that multiple individuals have the shortest path towards to all other individuals (in our case, direct connections) and that no individual is particularly distant from the rest of the network. Again, considering how connected our dataset is, this is only logical.

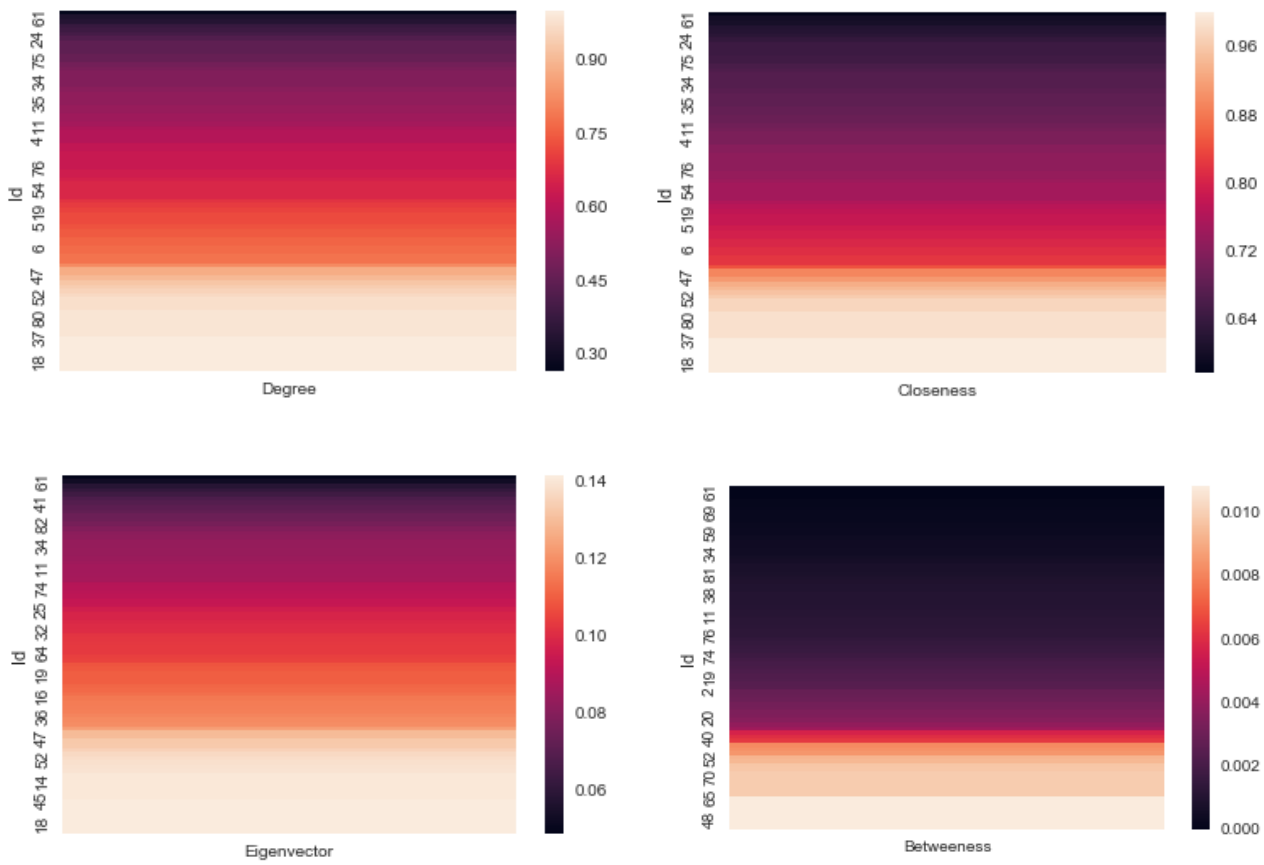

Figure 9 – Centrality measures for all members of the network

In terms of betweenness, the values given are extraordinarily low, with the highest value being 1% and a significant portion of the network having a value closer to 0%. Since this measure consists of the number of shortest paths that pass through a specific path divided by the total amount of shortest paths, these results can also be explained through the incredibly high amount of connections within the network. The high connectivity leads to a very high amount of shortest paths between all nodes in our network which results in low betweenness values. Finally the last metric that we want to comment, in terms of centrality, is the eigenvector. The existing eigenvector values are also relatively low, as they differ from about 5 to 14%. As mentioned in Chapter Literature Review, this metric is similar to degree centrality but takes into account the quality of the nodes each particular node has a connection with. We believe this occurs due to the fact that all nodes are highly connected, no particular nodes can be considered highly influential in the network meaning there are no particular connections to nodes with higher quality, which, in turn, leads to low values for this measure. Another important deduction is that, due to these very high values of connection, individuals that have connections to all members of the network, logically connect to those that have relatively lower amount of connections. Since these are considered as lesser influential individuals, individuals that have higher values of degree centrality ultimately have lower values of eigenvector centrality than one could expect.

4.1.1. Structures within the network

During Chapter Literature Review, we mention that one of the interesting ways to analyse a social network is to assess the different structures within it. This is because particular subsets of a network could possess specific details that are unlike any other part of the network. However, when attempting to assess these structures within our network we were not able to obtain relevant conclusions. This is related to the aforementioned characteristic of our network: the existing high connectivity between nodes. Since no individual is particularly poorly connected within the network, and there are a number of individuals connected to all individuals of the network, obtaining component subgraphs that stand out, be it by using an *Island method* or by identifying and isolating *Ego networks,* was not possible. In terms of triads, particularly those that are closed, the high number of connections throughout the network led to a very large number of these structures – a significant portion of the different combinations of nodes possible within the network. The number of cliques was also large and in terms of sizing could include a significant part of the network, leading to no further conclusions other than the high connectivity already mentioned. Hierarchical clustering was also used and, similar to the other techniques, it did not allow us to draw an interesting conclusion from it, as the high connectivity between the various individuals, particularly the high amount of shortest paths between nodes, did not lead to valid results. Finally, in terms of block models, these require the computation of a hierarchical clustering model (Tsvetovat & Kouznetsov, 2011). Since this was not possible, we could not apply this technique.

Despite the lack of conclusions we could take by using this kind of techniques, such as *ego* and *island* methods, its use was useful since it showed that, given this high connectivity that leads to a lack of possible structures within the network, the number of particular behaviours for a specific amount of individuals should either be small or even non-existent.

5. MUSICAL PREFERENCE HOMOGENIZATION

As referenced previously, one of our main goals in this work consists in assessing the homogenization of musical preferences within a social network. After doing an analysis of musical preferences' dataset, we are able to conclude that various changes in musical preferences do, in fact, occur over time. However, that does not necessarily mean that relationships between individuals have influence over these changes.

In order to assess whether musical preferences changes in terms of genres were influenced by existing relationships we take a look at differences between levels of preference within the various individuals that are connected through the network. Discussed in Chapter Dataset Analysis, each genre interest is classified as slight (represented as 1), moderate (represented as 2) and high (represented as 3) interest. This was done through creating binary variables in the social network dataset that represent whether individual A, who is part of our social network defined previously, has the same level interest as individual B, a connection of individual A within the network. This assessment will also be done from a yearly perspective, given the specific months with incredibly low amount of responses, as mentioned in Chapter Dataset Analysis.

In terms of data preparation, we obtained the relevant data by merging the musical preferences and relationships datasets and subtracting the level of preference of each genre of individual A by the level preference of individual B, followed by then using *if clauses* to make these differences in preference binary (1 if any difference exists, 0 if the interest is the same). For our second analysis, a similar method was done but only considering a change in interest when at least two levels of preference changes (ex: from 1 to 3 and vice-versa)

Taking into account the first approach, whose results are present in Table 5, the first conclusion that can be taken is that individuals have particularly different genre interests within the network, since, on average, these disagree with others that they are connected to about their genre interest around 70% of the time. This difference is particularly high for classic rock in 2008, where 90% of individuals with connections disagreed, and lowest for hip-hop / r&b, in which only around 50% of individuals with connections disagreed. These values might seem particularly high, but considering that all it takes for a difference in genre interest is an individual having high interest for a genre and another having only moderate interest for the same genre, these results do not seem illogical.

Genre	2008 (%)	2009 (%)	Difference (%)
indie / alternative rock	87,39	71,73	-15,66
classic rock	90,27	74,35	-15,92
classical	82,01	79,38	-2,63
techno / lounge / electronic	78,89	66,46	-12,44
jazz	72,93	67,23	-5,70
pop / top 40	78,86	72,17	-6,69
showtunes	77,22	70,81	-6,41
heavy metal / hardcore	69,19	62,00	-7,19
hip-hop / r&b	64,22	52,66	-11,55
country / folk	64,40	60,89	-3,51
other	81,89	65,68	-16,21
Avg	77,03	67,58	-9,45

Table 5 - Percentage of individuals with differences of at least a preference level of interest in genre interest per year

In terms of differences in levels of genre interest, there is a decrease when comparing 2008 and 2009. On average, these differences in terms of levels of interest consist of around 10%, with the genres indie/alternative rock, classic rock and other as the genres in which the level of interest seems to assimilate between individuals with connections by 15%. However, the previous analysis does impose a strict criteria, since individuals had to have the exact same level of interest, without even slight differences such as the ones between moderate or high interest. If we apply a more relaxed requirement, in which we consider only changes of at least two levels for an actual change of level of interest (the second approach mentioned), we can assess that results in terms of homogenization are very similar.

As seen in Table 6, in this assessment differences decreased around 9% on average between 2008 and 2009 among individuals that possess connections, with the same three as before (indie/alternative rock, classic rock and other) as the ones in which this decrease was larger. An important difference however is that, while in the previous analysis individuals that had connections disagreed between 50 to 90% of the time depending on the genre, using this criteria, people that possess relationships between each other disagree only between 15 to 50% of the time. This makes the 9% decrease on average an even more impressive evidence of homogenization, since this accounts to nearly a third of the existing differences between connected individuals on average in 2008.

Genre	2008 (%)	2009 (%)	Difference (%)
indie / alternative rock	39,68	25,46	-14,22
classic rock	48,83	34,90	-13,93
classical	47,00	37,80	-9,20
techno / lounge / electronic	29,20	17,91	-11,29
jazz	22,37	18,05	-4,32
pop / top 40	26,55	25,22	-1,33
showtunes	37,79	27,54	-10,25
heavy metal / hardcore	24,91	13,75	-11,16
hip-hop / r&b	14,40	4,89	-9,51
country / folk	18,26	15,30	-2,97
other	42,68	30,25	-12,42
Avg	31,97	22,82	-9,15

Table 6 - Percentage of individuals with differences of two preference levels or more in genre interest per year

Another interesting analysis we thought about consists in seeing these differences by type of relationship instead of over time. In order to quantify this, we applied the same criteria as in Table 5. Yet, instead of comparing 2008 with 2009, we differentiated each connection within the network using the previously mentioned level of relationship allocation (relationship by twitter or Facebook was quantified as a 1, socializing twice per week or being a political discussant were labelled as a 2, and being a close friend was qualified as a 3) and compared the difference between levels of genre interest between the three different types.

Similar to what we found in our analysis of Table 5, in Table 7 the percentage of individuals with links disagreeing is extremely high, ranging from between 58 to over 85%, depending on the genre and the level of interest. This is expected due to how rigorous the criteria is. However, when comparing genre interest differences between links with different levels of interest, there does not seem to be a consistent decrease of differences depending on how strongly an individual is connected to other. When comparing links with a relationship strength of 1 and 2, on average, connections with a strength of 2 only have lesser differences of around 1%. Furthermore, these differences do not seem consistent between genres, since, while decreasing somewhat significantly in classical and other genres (around 7 and 4%, respectively), they also increase 4% for classic rock, for example.

The same conclusions can be taken when comparing individuals with connections of strength 3 with ones of strength 2. On average, these differences actually increase in links with a level of 3 over ones with level 2 in about 1%, with certain genres having smaller differences in around 2% (other and heavy metal / hardcore) and others having larger differences, such as jazz and country/folk, whose difference increase in about 4 and 8%, respectively.

Genre	1 (%)	2 (%)	% of difference (2-1)	3 (%)	% of difference (3-2)
indie / alternative rock	78,89	80,04	1,14	79,82	-0,21
classic rock	79,52	83,51	4,00	84,12	0,60
classical	85,41	78,11	-7,30	78,62	0,51
techno/lounge/electronic	72,53	72,68	0,15	72,92	0,23
jazz	69,87	67,81	-2,05	75,46	7,64
pop / top 40	76,02	75,36	-0,66	74,73	-0,63
showtunes	75,63	72,81	-2,82	74,26	1,45
heavy metal / hardcore	64,93	66,59	1,66	64,66	-1,92
hip-hop / r&b	58,12	58,14	0,02	59,70	1,56
country / folk	61,91	61,72	-0,18	65,65	3,93
other	77,09	72,69	-4,40	71,13	-1,56
Avg	72,72	71,77	-0,95	72,83	1,05

Table 7 - Percentage of individuals with differences in genre preference/interest per relationship level

In conclusion of this chapter, two main insights can be taken from our analysis. The first is that there is an indication of homogenization of genre levels of interest throughout the network over time, particularly more significant the lesser strict we are in terms of levels of interest. The second is that the same cannot be said when comparing differences of genre interest preferences between different relationship levels. This means that we could not find any evidence that stronger connections between individuals can lead to a lesser amount of genre interest differences between them.

6. METHODOLOGY

One of the main purposes of the present work consists in assessing various changes in musical preferences over time to see if the existing social network is a relevant influence in these changes. This chapter explains the process behind this assessment, by describing the existing methodology for the model used, HMM (Hidden Markov Model), as well as providing a brief explanation of the model we will use for comparison purposes, SVM (Support Vector Machine). For HMM, the model definition that we provide below is an adaptation from the Rabiner (1989) work as well as other previous work that use HMMs which helped to detail our model (Da Silva, 2014;Daniel & Martin, 2018; Stamp, 2018). In terms of the SVM, the description given is inspired by the work of Suykens & Vandewalle (1999), Ben-Hur & Weston (2010) and Lee et al. (2012). In the present thesis I have described the model principles and how it was used to answer the thesis objectives but please refer to the papers mentioned above for more details about these models.

6.1. Hidden Markov Model (HMM)

HMMs date back to the late 1960s through various contributions given by Leonard Esau Baum (Da Silva, 2014). One example of contributions is from Baum & Petrie (1966) in which statistical inferences for probabilistic functions in Markov chains are described. Through an impactful tutorial created by Lawrence Rabiner (Rabiner, 1989) HMMs gained notoriety and started to be used in multiple speech recognition problems (Da Silva, 2014)

To fully comprehend HMMs, one should start by understanding how a more basic model functions, namely Markov chains. Within a Markov chain a system possesses multiple states S, comprised of N distinct states. At specific points in time, this type of system can have a change in a state or remain in the same state of the previous step. These state changes can be modelled with a state-transition matrix A:

$$A = \ a_{ij}, \forall \text{ i, j } \epsilon \text{ S; } \sum_{j} aij = 1 \, , i \, \epsilon \, S$$

Equation 5 - State-transition matrix for a Markov chain

In terms of the first state, it is represented through a probability vector π. In notation, a Markov chain consists of a 3 tuple $\varphi = (S, A, \pi)$. With this in mind, a n_{th} order Markov chain consists of a stochastic process through which the probability of assuming the next states is reliant exclusively on the last n states. Moreover, we also presume that transition probabilities are time-homogeneous. In other words, we assume these transition probabilities do not change based on the points in time that are being taken into account. For a first order Markov chain, the conditions below are in effect:

$$P(X_{t+1}|X_t = s_{t,} \, X_{t-1} = s_{t-1}, \dots \, , X_n = s_n) = P(X_{t+1} = i|X_{t-1} = j) \ \forall \text{ i, j}$$

Equation 6 - First condition fulfilled by a first order Markov chain

$$a_{ij} = P(X_{t+1} = i | X_t = j) = P(X_t = i | X_{t-1} = j), \forall \, t \in T, , \forall \, i, j \in S$$

Equation 7 - Second condition fulfilled by a first order Markov chain

Taking into account Equation 6 and Equation 7, we can now compute the probability of a sequence of states, S. Using the previous example of weather records consider the following state set and the transition matrix A:

$$S = \{S1, S2, S3\} = \{Hot, Mild, Cold\}$$

Equation 8 - State definition for Markov chain example

Transition Matrix A	Hot	Mild	Cold
Hot	0.55	0.15	0.30
Mild	0.25	0.45	0.30
Cold	0.30	0.30	0.40

Table 8 - Transition matrix definition for Markov chain example

Assuming a sequence of observations $O = \{S1, S1, S3, S1, S2, S2\}$, the likelihood that this sequence was obtained through the 3 tuple φ consists of the following:

$$\begin{aligned} P(O|\varphi) &= P(S_1, S_1, S_3, S_1, S_2, S_2|\varphi) \\ &= \pi * P(S_1|S_1) * P(S_3|S_1) * P(S_1|S_3) * P(S_2|S_1) * P(S_2|S_2) \\ &= 1 * 0.55 * 0.30 * 0.30 * 0.15 * 0.45 \sim 0.003 \end{aligned}$$

Now that we possess the relevant information regarding Markov chains, we can start to understand how HMMs work. The first significant difference is that while the existing states in a Markov chain were visible, in a HMM, S is now hidden. Moreover, each state has an observation O_k associated with it with a particular probability. These differences add an additional stochastic process. As such, HMMs are frequently defined as double stochastic processes. The observation set, O, as the name indicates, consists of what is visible and the emission of a particular observation at time t is reliant exclusively on the current state. The observation probability matrix, also known as the emission matrix, will be defined as B and the probability of O_k given S_j will consist of $b_j(k)$. With this in mind, we can now define HMMs as a 5 tuple $\lambda = \{\pi, S, A, O, B\}$.

Taking into account the definition of a HMM, as well as the mentioned information about Markov chains, in order to apply a HMM for our issue regarding musical preferences, and according to the existing literature mentioned previously, there are three particular questions that need to be answered:

1. Assuming a sequence of observations O, how can we calculate the probability that this sequence was computed by the model λ? Essentially, how can we calculate $P(O|\lambda)$?

2. Assuming a sequence of observations O, how can we calculate a state sequence S that, through some type of criteria, can be considered as the most optimal sequence?
3. Finally, how can we obtain the parameters for model λ that can lead to the maximum probability of $P(O|\lambda)$? In other words, how can we maximize the probability of O being emitted by the model λ?

In order to solve the first question posed, we need to first define $P(O|S,\lambda)$ for particular state sequence S:

$$P(O|S,\lambda) = b_{s_1}(O_1) * b_{s_2}(O_2) * \dots * b_{s_t}(O_t)$$

Equation 9 - Definition of probability of observation set know a state set and a HMM

Afterwards, similarly to what was done previously for Markov chains, we derive the probability of a particular state sequence: $P(S|\lambda) = \pi * a_{S1S2} * a_{S2S3} * \dots * a_{St-1St}$. As such, the probability of both O and S, $P(O,S|\lambda)$, can be defined simply as the product of two terms previously mentioned. With this in mind, in order to obtain $P(O|\lambda)$ we only need to iterate $P(O,S|\lambda)$ through all possible combinations of S. This can be defined in notation as such: $P(O|\lambda) = \sum_{all\ S} P(S|\lambda) * P(0|S,\lambda)$. In spite of this, if we took this simpler approach to iterate over all possible sequences of states, the computing complexity would be close to $2T * N^T$, where N consists of the number of states. This is something that is considered unfeasible even for smaller problems. There is, however, a much more efficient way to calculate $P(O|\lambda)$, known as the forward procedure. This technique consists of the following:

1. Initiation: $a_i(i) = \pi_i b_i(O_1), \forall\ i \in N$
2. Induction: $a_{t+1}(j) = \left[\sum_{i=1}^{N} a_t(i) * a_{ij}\right] b_j(O_{t+1}),\ j \in \{1,\dots,N\},\ t \in \{1,\dots,T-1\}$
3. Termination: $P(O|\lambda) = \sum_{i=1}^{N} a_T(i)$

During this process, we first define the forward-probabilities in function of both the initial state probability matrix and first observation given. In the second step, the expression that is represented between brackets can be summarized as the probability of observation sequence O, assuming that the current state is j at $t+1$. When we multiply the expression mentioned by $b_j(O_{t+1})$, we now possess the probability of having the emission O_{t+1} for state j. As such, a can be defined as the probability of the observation sequence O while being in state $(P(O,S_J|\lambda))$. Finally, we can compute $P(O|\lambda)$ by simply summing a over N. With this, we are now able to solve the first problem. In this part of this chapter, we will now explain how to define the backward variable β. While not necessary to resolve any of these problems directly, the answer to problem 3 does require the use of it. Since it is particularly similar to a, we believe that this part of our chapter is the correct time to introduce this concept. The backward probability, defined as $\beta_t(i)$, consists of the probability of a partial sequence starting at $t+1$ and ending at T, assuming model λ and that the state at time t is i.

$$\beta_t(i) = P(O_{t+1}O_{t+2}\dots O_T|q_t = S_i,\lambda)$$

Equation 10 - Notation for forward probability definition

The process that allows us to obtain this probability is as such:

1. Initiation: $\beta_T(i) = 1, \forall\ i \in N$
2. Induction: $\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(O_{t+1}) \beta_{t+1}(i), i \in \{1, \dots, N\}, t \in \{T-1, T-2, \dots, 1\}$

This process is essentially the inverse of the one used through the forward method, since this backward method starts at point T, the end of the sequence. All possible states j at $t+1$ must be accounted for, as well as the probability of observing O_{t+1} while in state j, in order to be in state j at time t while also taking in consideration the observation sequence that starts at time $t+1$. By also considering the remaining partial sequence from state j, as included in the definition of β, we reach what is considered above as step 2. As mentioned, we will leave the application of this concept to our explanation of solving problem 3.

Considering question 2 now, this is also known as the decoding issue. This is due to the fact that it essentially consists in obtaining the sequence of states are considered the best according to a particular rule. As such, we must define this rule. An example of a reasoning that could be used consists of choosing the state that emits a particular observation with the highest likelihood, under the logic of merely wishing to choose what is the most likely state individually at a specific discrete period of time. However, the most common criteria consists of selecting the sequence of states that has the highest probability of occurring when considering that a specific observation sequence O has occurred. A more formal definition of this consists in finding the maximum $P(S|O, \lambda)$ and can be obtained through using a dynamic programming technique named as the Viterbi algorithm (Viterbi, 1967). In order to have a better comprehend this algorithm we begin by defining δ:

$$\delta_t(i) = max_{S_1, S_2, .., S_{t-1}} P(s_1 s_2 \dots s_t = i, O_1 O_2 \dots O_t | \lambda)$$

Equation 11 - Notation for highest probability score through the Viterbi algorithm

With the Equation 11 in mind, we can define the variable δ as the highest probability score for a specific state S and a partial observation O ending at time t and state i. Through induction, we can provide the following definition: $\delta_{t+1}(j) = \left[max_i \delta_t(i) * a_{ij}\right] * b_j(O_{t+1})$

According to our criteria, in order to find the optimal sequence, we must save records of the states that, for each t and j, maximize N. This log is will be saved in array $\psi_t(i)$, as shown in the procedure for this below:

1. Initialization:

   $\delta_t(i) = \pi_i b_i(O_i),\ 1 \leq i \leq N$

   $\psi_1(i) = 0$

2. Recursion:

   $\delta_t(j) = max_{1 \leq i \leq N}\left[\delta_{t-1}(i) * a_{ij}\right] b_j(O_t),\ 2 \leq t \leq T, 1 \leq j \leq N$

   $\psi_1(j) = argmax_{1 \leq i \leq N}\left[\delta_{t-1}(i) * a_{ij}\right],\ 2 \leq t \leq T, 1 \leq j \leq N$

3. Termination:

$$P^* = max_{1 \le i \le N}[\delta_t(i)]$$

$$S_t = argmax_{1 \le i \le N}[\delta_t(i)]$$

4. Backtracking:

$$S_t^* = \psi_{t+1}(S_{t+1}^*), t = T-1, T-2, \dots, 1$$

This algorithm, with the exception of the backtracking part, has a similar behaviour to the forward procedure mentioned previously. The only exception is that only the maximum value is saved, avoiding the computing complexity mentioned previously. The Viterbi algorithm's behaviour consists in obtaining the optimal state sequence when taking into account an observation sequence O by memorizing the argument that provides the maximum probability along a single path, for each j and t.

The final issue we need to tackle is question 3, considered as the most difficult one. This consists in understanding how we can tune the parameters for the model λ in order to improve the likelihood of an observation sequence O was produced by λ as much as possible or, in notation, $P(O|\lambda)$.

The difficulty of this problem is due to two main factors. The first consists of the fact that, since our states are hidden, we cannot define our parameters by using the actual probabilities of observations occurring considering each hidden state for our emission matrix, nor can we use the actual probabilities of each state given the state that occurred at a previous time, also known as our previously mentioned transition matrix. The other factor is that there is no exact possible way of estimating best model parameters i.e. we can only obtain local maximum values.

The method that will be used to tackle this type of issues consists of a process known as Baum-Welch (Da Silva, 2014) that only stops attempting to get the best parameters when either a previously given number of iterations is reached or the increase on the probability of a certain sequence occurring is marginal. This process consists of a use of the forward and backward variables, calculated through the previously mentioned methods.

First off, we begin by considering $\xi_t(i,j)$ as the probability of a state i occurring at time t and state j occurring at time $t+1$, assuming a sequence of observations O and model λ.

$$\xi_t(i,j) = P(s_t = S_t, s_{t+1} = S_j | O, \lambda)$$

Equation 12 - Notation of probability that two particular states occurred at two different points of time, given the sequence of observations and the existing HMM model

With this definition in mind, and when taking account the Bayes rule, which says that the probability of event A given event B occurring can be calculated by dividing the product of the probability of A occurring and the probability of B occurring given A by the probability of B occurring, as well as the information previously obtained of the forward and backward algorithms, the aforementioned variable can be defined as such:

$$\xi_t(i,j) = \frac{P(q_t = S_i, q_{t+1} = S_j, O|\ \lambda)}{P(O|\lambda)} = \frac{a_t(i)a_{ij}b_j(O_{t+1})\beta_{t+1}(j)}{P(O|\lambda)}$$

Equation 13 - Developed notation of Equation 12 based on the Bayes rule and the forward and backwards algorithms

in which $a_t(i)$ represents part of the existing sequences, up until t, $b_j(O_{t+1})\beta_{t+1}(j)$ represents the change to the state j and the observation O_{t+1} ,taking into account that the current state is j and $\beta_{t+1}(j)$ represents the part of the existing sequences that go from $t+1$ until T. As such, it is logical to comprehend that, through aggregating ξ_t over $T-1$, we can obtain the assumed number of transitions from state i to state j. This sum only occurs up until the second to last discrete time period ($T-1$) due to the fact that, as one can assume, there is no transition at the last discrete period (T). The next step consists of defining the expected number of transitions from state i, using the previous inferences. To do this, we start by defining the probability of being in state i at time t given an observation sequence O:

$$\gamma_t(i) = \frac{\alpha_t(i)\beta_t(i)}{P(O|\lambda)}$$

Equation 14 - Notation of probability of a state occurring at specific time given an observation sequence

By aggregating γ_t through -1 , we can obtain the expected number of transitions from state. As such, with all of this information, we can now introduce the following definitions:

$$\sum_{t=1}^{T-1} \gamma_t(i) = Expected\ number\ of\ transitions\ from\ state\ i$$

Equation 15 - Expected number of transitions from a specific state

$$\sum_{t=1}^{T-1} \xi_t(i,j) = Expected\ number\ of\ transitions\ from\ state\ i\ to\ state\ j$$

Equation 16 - Expected number of transitions from a specific state to another

For the last part of this process, we can use these two previous formulas to obtain the values for the initial probabilities, the transition and the observation matrixes. These can be represented as follows:

$$\pi_i = \gamma_1(i),\ 1 \leq i \leq N$$

Equation 17 - Initial probability calculation based on the Baum-Welch procedure

$$a_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t}^{T-1} \gamma_t(i)}$$

Equation 18 – Transition probability calculation based on the Baum-Welch procedure

$$b_j(k) = \frac{\sum_{t_{s.to}\ O_t=k}^{T-1} \gamma_t(j)}{\sum_{t}^{T} \gamma_t(j)}$$

Equation 19– Emission probability calculation based on the Baum-Welch procedure

Now we have all the information we need to calculate the HMM parameters through the Baum-Welch algorithm for our problem.

6.2. Support Vector Machine (SVM)

In our work, we use a HMM to assess whether we can infer change of musical preferences over time considering the preferences of an individual as well as others within the network. However, in order to present a good evaluation of this model's strengths and weaknesses we used a different model with some similarities to our HMM that could provide an interesting baseline. In this part of our work we provide a brief explanation of what this algorithm consists, as well as of some of the parameters that can be used to tune it. We also provide our reasoning in terms of why we believe the results of this model can be used for comparison.

In a SVM data is mapped into a higher dimensional input space where an optimal separating hyperplane is constructed, whose one of its applications consists of resolving classification problems (Suykens & Vandewalle, 1999). This hyperplane can be defined through either linear or non-linear methods. While a linear method directly defines a hyperspace whose dimensionality depends exclusively on the number of variables used, in nonlinear methods this depends on what is known as the *kernel trick* (Lee et al., 2012). In the latter, variables are mapped to a specific higher dimensional plane, which varies depending on the particular kernel function that is applied in order to find a hyperplane capable of separating data that is impossible to linearly differentiate.

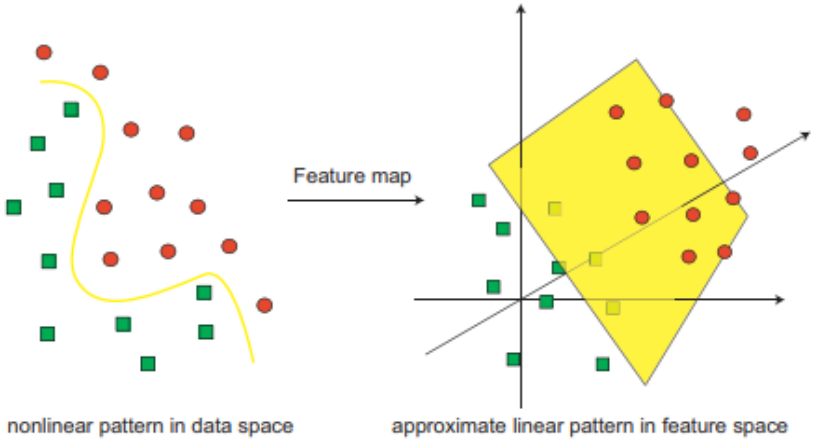

Figure 10 - Illustration of a nonlinear SVM (Lee et al., 2012)

In terms of parameters that can be used to tune our SVM, one of these consists in what is known as gamma. It is a hyper parameter that assesses just how sensitive the model is to the influence of each training point. In terms of the hyperspace defined, gamma will impact just how curved this decision boundary will be (Ben-Hur & Weston, 2010). If it is too low the hyperplane will be straighter approaching an almost linear version. If it is too high it will lead to this decision boundary being so curved it will be specifically detailed to the training data given, essentially causing the model to be overfitted.

The other two parameters consist in C and degree. C, also referred as the soft margin constant, is a penalty parameter that reflects just how critical the model is on values close to the margin of the hyperplane (Ben-Hur & Weston, 2010). In the hyperplane, this value mainly changes on just how wide the margin of this decision boundary will be. When this penalty value is very low, data points close to the boundary are ignored and the margin is bigger. If

the value is very high, the margin becomes very slim and, in an attempt to separate data points with slim margin, its orientation also changes.

In terms of the degree this parameter is exclusive to the polynomial function since, as the name indicates, it represents the degree of the function itself, with a degree of 1 being the equivalent of a linear function. This has a similar affect to the hyperplane as the gamma coefficient mentioned, since it increases the curving of the decision boundary, as well as leading to overfitting (Ben-Hur & Weston, 2010).

Concluding this chapter, we believe this model will provide a relevant comparison due to its similarities to our HMM, since it also uses existing characteristics of the data used to train it to define whether or not a different characteristic exists within the network. In a HMM, the existing observations are used to assess whether a hidden state occurred based on probabilities while in SVM the variables given are used to define a hyperplane to assess whether the dependent variable is positive or not. The main difference is that while SVM predicts every data point separately, the HMM, through the Viterbi algorithm, predicts sequences based on the established discrete points in time. Another reason is the fact that there has been precedent in using SVM as a comparison to a HMM for prediction, particularly in terms of assessing the spread of infection across a social network (Dong et al., 2012).

7. MODEL IMPLEMENTATION

As you can understand from the Chapter Methodology a HMM requires that we define specific concepts. In this chapter we begin by defining our observations and hidden states, as well as what will consist of a discrete period of time. We then explain how these concepts impact the preparation of our data, both for HMM as well as SVM. Finally, having in mind the parameters used we detail how our test dataset was prepared.

Let's begin with the critical part of our concept definition: representing our discrete states of time. As discussed in Chapter Dataset Analysis, the musical preferences survey was taken multiple times over several months with three specific months that have a very low amount of responses. Hence, if we take our discrete states of time as monthly snapshots, each different month would be a different point of time, consequently leading to data from these three months essentially being disconnected from the rest. A yearly snapshot would also not make much sense since the majority of the individuals responded multiple times per year. As such, we chose to assume our discrete points of time using the frequency of survey response. This means that our first point of time includes the answers that each individual responded the first time they answered the survey, the second one possesses the answers from the second time, and so on and so forth. This means that we no longer rely on particular snapshots that might exclude answers while still minimizing discrepancies as much as possible in terms of response timing since the vast majority of individuals respond for the first and second time around September/October and for the third/forth time around May/June.

In terms of our hidden states, since we wish to assess if an individual changed preferences over time these states must consist of a variable that represents that an individual had a significant change from one discrete time period to another. When taking into account that individuals have interest on average around 6 genres, through the experimentation that we will analyse later on, it was decided that we will consider our hidden state a binary variable named *change* – that consists in whether an individual's preferences at a particular point of time are different from the preferences in the following point of time, in at least 4 genres. When we are referring to preferences, we are referring to binary variables that only represent whether an individual has an interest or not in a specific genre, without taking into account the level of interest of an individual in each genre. The main assumption here is that this change of preferences between a current state and the following state always occurs at the current state, with the following state only being used due to being the only way to assess whether change occurred in the current state. This allocation of *change* to the current state rather than the following state will be something that will also be discussed during the Chapter Experimentation and Discussion.

The last concept we need to represent consists of what is defined as an observation. As explained in Chapter Methodology, at a particular state of time, a specific state emits a single observation. However, in order to infer change of preferences, we wish to use data regarding each individuals' genre preferences, as well as the preferences from the network. With this in mind, our observations will consist of a combination of both the individuals' preferences and the preferences of the individuals that each specific person is connected

through the network. More specifically, these consist of a combination of each individual's own preferences in all 11 genres as well as a *change of individual B* variable. In order to better explain this variable, an example is warranted. Let's say individual A and individual B are two different people who are part of our network and are connected. In our data, for individual A, one of the existing observations that will exist consists of a combination of the 11 genre preference binary variables, representing whether individual A had any kind of interest at a specific point of time, as well as a 12^{th} variable, that represents whether individual B changed preferences at that particular point in time, taking into account that *change* possesses the same definition as the one given for our hidden states. However, as we assessed through our network analysis, an individual has multiple connections within the network. This means that, in our data, each individual is represented multiple times for a specific point in time in our data, according to the amount of connections each individual has within the social network. This can be seen through Table 9, in which the combination shown only changes between lines due to the fact the two individuals connected to individual 1 had different *change* variables associated to them. The reasoning as to why we chose these specific variables for the combinations of our variables will be further discussed in Chapter Experimentation and Discussion.

Id A	Id B	Response number	Combination	Change
1	51	1	011110011100	1
1	32	1	011110011101	0

Table 9 - Example of the final representation of our data for the HMM

Since, as mentioned in the previous paragraph, we have multiple data points for the same individual in a given point of time, we will have two different types of results, one in which we take into account predictions per data input and another where we take the predictions given for an individual at a specific point in time. Finally, another important detail to mention regarding our HMM is the fact that it consists in a first order HMM, whose main premise is that a state in the next point in time depends exclusively on the current state.

For SVM, we will use the variables used for defining our combination in the HMM as our input and the *change* variable represented as our hidden state as the dependent variable. By using essentially the same variables, we will be able to draw a fair comparison between these models.

For our test and training sets, these were separated through a 90/10 split, in which all data points regarding each individual A were kept in the same set (ex: all rows for id A 31 are in our training set, while all rows for id A 10 are in the test set). As such, for our test dataset, we randomly selected ids of individuals until the data points regarding these made up 10 % of our complete dataset. This definition of 10% was due to both the issues with unique combinations described during the experimentation part of our work, as well as the fact that there has been precedent in using this percentage of population for predicting using an HMM, specifically for predicting spread of infection across a social network (Dong et al., 2012).

In terms of the decision to keep data regarding each individual A in the same set, this was done due to the fact that the HMM predictions take into account variations between each discrete point of time. As such, these predictions cannot be per individual line in our dataset.

Instead, we take all test data regarding particular individuals, chosen at random, in order to apply the Viterbi algorithm for all sequences pertaining to particular combinations of individuals' A and B, obtaining the likeliest sequence for each combination. We use the same exact test dataset for our SVM, in order to provide a fair comparison.

The last definition needed to obtain our results consists in the parameters for both SVM and HMM. In the HMM, the initial probabilities, as well as the emission and transition matrixes, are initialized randomly, since this is the most common procedure in the existing literature (Da Silva, 2014). We will then apply the previously mentioned Baum-Welch algorithm using our existing data, in order to possess the probabilities that will be fed to the Viterbi algorithm for predicting the most likely sequence. This algorithm was applied using 10,000 iterations. For our SVM, we will use a radial basis function (also known as RBF) kernel with a gamma of 0.083 (value is 1/12, rounded to thousandths is 0.083) and a C parameter of 10. Further details on how the iterations in the HMM and the parameters for SVM were chosen will be given in the Experimentation and Discussion part of our work.

8. RESULTS

Now that we understand how both of our models will be implemented, in this chapter we will analyse the various results we obtained through these. We can divide this chapter into two parts. The first one consists of introducing performance metrics that will be used to assess our model's performance. This was done by using the test dataset to predict this population's different hidden states using the Viterbi algorithm and comparing these results to the actual hidden states that occurred. The metrics consist in Accuracy, Precision, Recall and the F-measure, whose mathematical formulation are described as follows (Hossin & Sulaiman, 2015):

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

Equation 20 - Formula for obtaining the accuracy metric

$$Precision = \frac{TP}{TP + FP}$$

Equation 21 - Formula for obtaining the precision metric

$$Recall = \frac{TP}{TP + FN}$$

Equation 22 - Formula for obtaining the recall metric

$$F - Measure = \frac{2 * Precision * Recall}{Precision + Recall}$$

Equation 23 - Formula for obtaining the f-measure metric

TP – A true positive, in which the actual classification equals 1 and predicted classification was 1
FP – A false positive, in which the actual classification equals 0 but the predicted classification was 1
FN – A false negative, in which the actual classification equals 1 but the predicted classification was 0
TN – A true negative, in which the actual classification equals 0 and predicted classification was 0

Accuracy measures how many correct predictions exist when considering the total number of predictions, precision measures how many positive data points were correctly predicted when considering all positive predictions, recall measures the amount of actual positive data points that were correctly predicted, and the f-measure consists of a "harmonic mean of recall and precision". F-measure is generally considered as the better overall evaluation metric from the ones mentioned (Hossin & Sulaiman, 2015). As such, both in our analysis as well as in our experimentation this was the metric of reference.

As indicated in Chapter Model Implementation, these metrics will be analysed in two different ways. First off, we will take a look at these in a standard way, meaning we will calculate these metrics by comparing the predicted output of the Viterbi algorithm with the true values of the *change* dependent variable. However, as detailed previously, for a single individual A at a particular time, our dataset possesses multiple inputs of data, one for each individual that A has a connection with, despite the fact that individual A, at a point in time, has either changed preferences or not.

The fact that multiple rows for an individual at a specific point in time exist means that we have the same value of a dependent variable for multiple inputs (ex: for id 31 in a given point in time, we have multiple rows indicating that he changed preferences, one for each connection he has within the network. As such, our second way of assessing the metrics previously mentioned will consist in assessing if, for all predictions made for an individual A in a single point in time, at least 50% of these are of *change*. If so, we consider the prediction for this individual at this time as *change*. If it is lower than 50%, we consider *no change*. This allows for a more accurate representation of what we are trying to predict, since it no longer considers multiple inputs for each individual at a point in time, but rather a single one for each.

The second part of this chapter will consist of an analysis of both the transition matrix used and the probabilities of *change* based on each of our variables that was used in obtaining the combinations that consisted of our observations for our HMM. This is due to the fact that, since these observations are a combination of the individuals' preferences, as well as the *change of individual B* variable, a simple look at the emission matrix would only showcase the probabilities of *change* or *no change* based on the combinations themselves, which doesn't provide much insight. As such, we will use the probabilities of the emission matrix to calculate the probabilities of *change* based on each of the individual's preferences and the *change of individual B*, by summing all probabilities of combinations that consist of the probability of *change* or *no change* knowing that the individual had a particular preference or that individual B changed preferences. For example, for the variable *change of individual B*, we will sum all the probabilities of combinations that have this variable as 1 considering there was change, and then do the same but considering there was no change.

In Table 10 results for each of our models are shown. Our HMM performs reasonably well, with all metrics described being around 65 to 75%, both in our line by line predictions as well as the in the predictions by id. The SVM mentioned has also given positive results with all metrics also in the 70 to 80 % range.

Metric	HMM	SVM	HMM (per id A and time)	SVM (per id A and time)
Accuracy	0.70	0.85	0.64	0.81
Precision	0.72	0.86	0.78	0.82
Recall	0.71	0.85	0.64	0.81
F-Measure	0.71	0.85	0.70	0.81

Table 10 - Prediction metrics results

In terms of assessing these initial results, we find that both types of prediction give slightly different results. While the SVM clearly has better results with values ranging between 0.8 and 0.9, the HMM only possesses values between 0.65 to 0.75 in accuracy, precision and f-measure, and a value of 0.64 in recall and f-measure per id A and timeframe. This disparity between results in different metrics for our HMM might be due to the model seemingly trending towards having a higher proportion of predicted changes that were correct (precision) than the proportion of actual changes that were classified correctly (recall).

In terms of why the differences in performance between two models might be occurring, we must take into account that, when considering the examples mentioned in Chapter Literature Review, HMMs seem to be typically used with larger amounts of data than the musical preferences dataset described previously. This is true, for example, in one of our most influential papers that studied the spread of infection within a social network through a version of an HMM and also compared to an SVM (Dong et al., 2012). In this paper, data used was obtained daily instead of only specific monthly snapshots. As such, we believe that the lack of data is the reason for the HMM to have worse performances, particularly since it caused several limitations that will be further discussed during the Experimentation and Discussion part of our work. When comparing the two different types of predictions, the standard version showcases slightly better performance levels than the predictions per individual at a certain point in time, which means there is an indication that our model, when considering individuals at specific points in time, does predict worse. However, given that the difference is relatively small, no additional conclusions can be taken, particularly given that due to the group by effect the number of actual predictions decreased significantly.

Overall, both models' performances are relatively positive, implying that these do have some prediction capability. In other words, there are some indications of both an individual's own musical genre preferences and whether individual B changed preferences or not influencing change of musical preferences within our network. Despite the fact that the SVM performed slightly better, we believe that the HMM is still the better model for two reasons. First off, from a logical point of view, and in our opinion, it makes more sense to consider musical preference change as something that occurs over time, rather than assessing this for each point in time and each individual independently. This is particularly true when taking into account we are also are considering the influence of a social network in these changes. As such, since the SVM predicts for each point in time independently, while the HMM makes its predictions considering variation over discrete points in time, it is logical to consider that the HMM provides a more faithful representation of how musical preferences change. Secondly, from a more practical point of view, the fact that the HMM predicts in this sequential manner is crucial because it gives us more information regarding the behaviour over time of the population in terms of change of musical preferences. An example of this is the fact that, through the Baum-Welch algorithm, the HMM computes an emission matrix, that estimates the probability of whether individuals that changed in the previous state will change again in the current state, which is discussed in the next couple of paragraphs and represented in Table 11. Since our SVM does not provide this type of information, practically, we also cannot deem it a better model, despite the slightly better performance in the measures assessed.

Focusing now exclusively on the HMM, here is the transition matrix for our hidden states (change/no change)

Hidden state	Change (current state)	No Change (current state)
Change (previous state)	0.61	0.39
No Change (previous state)	0.40	0.60

Table 11 - Emission matrix for our Hidden Markov Model

As you can see from Table 11 when an individual changed musical preferences in the previous state, the individual tends to also change in the following state. This same conclusion applies to *no change*. This indicates that there are individuals that tend to consistently change preferences (likely just enjoy experimenting new genres), as well as the possibility that there are other individuals that tends to continually enjoy the same genres over time. However, since the split between these probabilities is of only around 60/40, this tendency is only slight.

In this next analysis, we will take a look at the estimated probabilities of change or no change from our model, calculated by aggregating the probabilities for each combination that possessed each particular variable as positive (obtained through the emission matrix), depending on if they occurred along with change or no change.

Variable	No Change	Change
Classical	0.72	0.70
Classic Rock	0.59	0.64
Country / Folk	0.39	0.44
Heavy metal / Hardcore	0.30	0.42
Hip-hop / R&B	0.31	0.38
Indie / Alternative Rock	0.59	0.70
Jazz	0.61	0.56
Other	0.39	0.47
Pop / Top 40	0.48	0.54
Showtunes	0.36	0.66
Techno/ Lounge / Electronic	0.48	0.69
Change Individual B	0.33	0.38

Table 12 - Probabilities of change/no change based on the variables present in the observations of our model being positive

As you can infer from Table 12, most variables tend to cause a higher probability of *change*, rather than *no change* in our model, with the exception of Classical and Jazz musical preferences. This, of course, includes the variable that consists of *change of individual B* that is connected to the network. The most impactful variables for inducing change within our HMM model are Showtunes and Techno / Lounge / Electronic, which have a difference of 20 to 30 % between the probabilities of *change* and *no change*, likely indicating that individuals that enjoy these genres change their preferences in other genres over time more frequently. However, there is a significant discrepancy between the amounts of data that exist in which each flag variable is positive. For example, while individuals in our training data possess an interest in classical genre around 70% of the time (both probabilities are of around 70%),

interest in Hip-hop / R&B only occurs around 30 to 40 % of the time. In terms of our *change of individual B* variable, individuals of the network seem to change preferences only around 30 to 40 % of the time. This might mean that this variable isn't that impactful in our model.

9. EXPERIMENTATION AND DISCUSSION

In this part of our work, we will start by showcasing our main thought process, as well as the various experiments during the creation of our model, including the assumptions made to surpass the difficulties we found. Afterwards, we will take a deeper dive into the results mentioned in the previous chapter, contextualizing these with the experiments and assumptions made.

First off, the main goal at the start of the development of our model consisted in measuring the influence of a social network in an individual's musical preferences. Our first thought was of a model that could allow us to see the impact of preferences of individuals of network in the preferences of a different individual that was connected in the network. This posed our first question in terms of defining our model. Furthermore, we wished to take into account the evolution of these preferences over time, in order to see if there was propagation of these preferences. With this in mind, and taking into account the methodology of a HMM mentioned above, we concluded that this model was the most logical option since it fits the requirements mentioned. As such, and now that we understood what model made the most sense for our problem, we began defining all the necessary logic for it. This mainly consisted in the three different definitions already mentioned in Chapter Model Implementation, which consist of our discrete points in time, our states and our observations. In terms of defining our points in time, the decision was relatively straightforward, as it was already explained during our model definition. For our states and observations however, the definition of these took multiple experiments, which we will now detail.

In terms of what are our states, as mentioned in our model definition, these must be hidden, meaning that they should consist of variables that have an unknown root cause. Given that we wish to better understand an influence in preferences, our hidden states would have to consist of existence of *change*, or lack thereof, within these. This did, however, impose two important limitations, as well as a particular concern. The first one consisted of the fact that, for calculating *change*, we would have to compare the preferences in a particular state with either the state before or state after to see if there were any changes in preferences. This meant that either our first state and last state for each individual would have to be used exclusively for this definition, since, because we only have a limited amount of states, we would not be able to assess whether or not it changed based on the previous or following state. Since, on average, an individual answered between 3 to 4 surveys, this implicated a loss of a significant portion of the total amount of rows in our dataset. It would also implicate an important assumption. As mentioned before, a future state in a HMM can only depend on the present state. As such, we would have to assume either a change between states x and x+1 occurred in either the second state or the first state. Ultimately, as explained in our model definition, we chose to assume that a change between state x and x+1 occurred in the first state, since, from a logical point of view, it made more sense to say the individual changed preference in state x but we could only assess whether this happened or not when taking into account the preferences in state x+1. The second limitation consisted of the exact definition of *change*. In other words, just how many changes in either genre preferences (binary approach) or levels of preference (takes into account interest) are required for our hidden

state to be considered as *change*. Two important factors were in play in this decision. The first is the fact that, as referred in Chapter Dataset Analysis, each individual at a particular time has interest in around 6 different genres. This fact right away limited our search for a binary definition, since it would seem less logical to assume *change* by a change of 6 or more binary genre preferences (implies that from a point in time to another, an average individual would have to dramatically change his interests). The second fact consisted in that, after assessing our dataset using our discrete points in time, we found a constant fluctuation of at least one or two musical genre preferences and multiple genre preference levels per individual per survey, making it so, for example, defining *change* as just the change in the musical preference of a single genre would be impossible since the dataset would be overwhelmingly skewed towards positive values of *change*. Another point of view that lead to our *change* definition was that, for an individual to truly change musical preferences, a simple slight change in terms of a new genre that he is listening to that is new or a slight increase in interest of change of genre the individual is already interested in would not be a meaningful musical preference change from a logical point of view. This lead us to test multiple hypothesis for defining what is *change*, by changing the amount of either changes in levels of interest in the different genres or changes of interest in the genres themselves. At this point, various tests were done to provide a dependent variable that is both balanced and provided the best results (measured through the metrics present in the results part of our work) in the various HMMs that we were experimenting with (later described since the difference between these consisted in the definition of our observations). After this analysis, we found that assuming *change* as either 4 genre preference changes or 6 changes in levels of interest in the different genres, depending on whether we would later on use binary variables of preference change or levels of interest in each variable, provided the best combination of a balanced dataset and better results for our model, each having results of 0.65 to 0.75 in the metrics analysed, as opposed to other definitions (ex: 5 genre preference changes, 7 levels of genre interest, etc. that barely had any results that surpassed 0.5 in any metric). This also seemed to make sense when considering that our dataset possesses 11 genres, so assuming *change* as an individual changing preferences in 4 genres or changing the level of preference in 6 seemed like a meaningful change, but not overwhelmingly so.

Concluding our experimentation for the definition of our states, one concern we had was the fact that, due to calculating our states based on the existing preferences (which will be part of our observations), the correlation between these would cause doubt in the validity of our results. However, since these states are considered hidden, and as you can infer from Chapter Methodology, a HMM works through unsupervised learning, meaning that it doesn't take into account a dependent variable, merely attempting to find patterns according to the data that is given, we don't believe that the fact that this state is calculated in the fashion previously mentioned impacts the validity of inferences. This might seem counter-intuitive since typically unsupervised learning algorithms aren't used in classification problems, but there has been precedent in assessing a type of HMM in such a manner (Dong et al., 2012)

In terms of our observations, as the name indicates, a HMM considers these as what was observed at a point in time. In simpler models, an observation is considered one of the existing distinct possibilities for a single variable. However, in our model, we wish to take into account multiple variables that occur with a specific value each at every possible state. As such, our observable states could not be defined as each one of our genre preferences or

levels of preference, but rather as combinations of these (similarly to the example given in Chapter Literature Review, where each observation could be considered as a combination of clothing items). For example, a possible observable state for individual A, if we are using only binary genre preferences from this particular individual, would be of him having interest in all 11 possible musical genres at that particular time.

This particular characteristic, combined with our existing limited dataset using survey data, immediately reduces our possible definitions. Ideally, we would have liked that each prediction was upon the change in preferences of a particular individual at particular point in time, using both this individuals' preferences and the preferences of all individuals in the network that he has a relationship with. However, as discussed in our network analysis, our network is highly connected, with multiple individuals being connected to our entire network. This fact, combined with only having a maximum of 4 answers to the musical preferences' survey for a particular dataset lead to a dataset of the type mentioned having very specific combinations for each individual in each state, that weren't be replicated by any other individual in any other state. This meant that, when taking out part of the network's population for our test dataset, these would include combinations that were unique and thus our model would be unable to predict them. This combination problem would also prove true even in simpler versions of our model. Even when our combinations have only preferences of an individual A and preferences of an individual B that has A has a connection to, already making it so we would have to predict multiple values for a single value of our predicted variable, each combination was also rare enough so that we could not obtain a relevant population for testing without any unique combination. However, since we were closer to unique combinations using this style of a dataset, we kept this logic but started attempting different types of simplifications. This was also the point that we decided we would also attempt to measure the various prediction metrics not only using the standard way but also grouped by individual A and time. This issue was also true when considering levels of preference. Since these possess a wider range of possible distinct values for each variable (values can vary from 0 to 3), the observations generated lead to an even larger amount of possible combinations. Even when attempting to group these possible values, by either grouping moderate and high interest (2s and 3s) levels or creating binary variables that consist of whether an individual has at least moderate interest in a genre or not, we could still not obtain a usable test dataset.

Since we were unable to obtain test data with non-unique combinations using the datasets mentioned in the previous paragraph, we found that we would have significantly reduce our variables. The best possible way we found to do this was to reduce either the preferences of individual whose preferences we are trying to predict (individual A), or the individual who A has a relationship with (individual B). However, this feature reduction could not increase the amount of distinct possible values in each feature, since this inevitably would also lead to too many unique combinations. We also did not wish to remove any specific genre from taking part in our dataset, given just how balanced our dataset is, as well as the relatively low correlation between each genre, which we detailed during Chapter Dataset Analysis. As such, we found that the best solution would be to, instead of including all preferences of individual B, include a single variable that consists of change of preferences of this individual. This makes logical sense since this variable would essentially represent whether an individual B changing preferences leads to individual A also changing preferences or, in simpler terms,

the influence of individual A's social network in his change of musical preferences. It also poses no issue in terms of validity since, when calculating our hidden state (*change*) only the preferences of individual A are taken into account. This variable of *change of individual B* was tested using two main definitions. These consist of whether the change mentioned compares the preference variables of the current state with the ones from the following one, or it compares the current state of these preferences with the previous one. These, of course, would be calculated using the same definition given for our hidden state (which would end up being defined as changing at least 4 genre preferences). Since the first definition mentioned (using current state and the next one) provided the best results (the ones mentioned in the previous chapter) and is the one most consistent with the state allocation mentioned when defining our hidden state (changes between state x and x+1 occur in the first state mentioned), we chose to go with that option.

Some other interesting variations of the model described above were also tested to see if they provided interesting results. The first one of these consisted in essentially multiplying the *change of individual B* variable by the relationship level between individual A and individual B, within the network. This test considered our *change* and *change of individual B* variables using change of 4 genre preferences. Unfortunately, it provided too many unique combinations for us to obtain a decent test dataset.

In terms of using levels of preference in genres for our observations, due to the already mentioned issues with obtaining unique combinations, testing using datasets that included these variables was almost impossible, since in almost every dataset attempted there were too many unique combinations for us to obtain a proper test dataset. In fact, the only dataset we could test consisted of including levels of preference of individual B (member of individual A's network) and at most a specific variable that represented if individual A had changed preferences when comparing preferences from the previous state with the ones from the state before. However, not only do these definitions of observable states give worse results (all metrics used were at best between 0.5 and 0.6), the variable that represents individual A takes into account previous states, putting into question our results, since our model is supposed to be a 1st order HMM (each current state only depends on the previous state). With this in mind, we decided not to use levels of preferences for our model, also leading to us discarding the possibility of using change in 6 levels of preference as the definition of our hidden state.

Another experiment that was attempted consisted of only using the change of preferences of individual B used in the defining model, in order to assess if the influence of the network alone could actually predict change of an individual in an accurate form. The values that we obtained were of around 0.5 in all of the metrics referred which fell short of a model we believe we could use for any kind of inference.

Now that we have explained the various experiments used for defining our main concepts, we will now detail how we defined the parameters for both the HMM and SVM. In terms of our HMM, the only parameter we used to tune consists of the number of iterations of the Baum-Welch algorithm. Due to the data limitations already mentioned, establishing a smaller number of iterations, such as 1 or 10, would lead to wildly different results in the predictions through the Viterbi algorithm (in multiple different executions of our model, the

standard deviation of the results in all metrics were up to 16%), depending on the random values given during the initialization of our procedure. In order to minimize this variance, we increased the number of iterations to 10000, since, with this number, the standard deviations mentioned previously dropped to below 2% for each metric. This was, in fact, the only path we found to obtain consistent values for our prediction metrics for this model. In terms of the SVM parameters used, one of the parameters we needed to define is the kernel function used. To do this, we ran this model using a linear function, as well as using polynomial, RBF and sigmoid functions. Ultimately, the SVM with a RBF function proved the one with the highest performance according to the metrics analysed (0.8 to 0.9 in all metrics when compared to the other kernel functions that gave at best 0.70 across all metrics). Therefore, we chose to go with this function. In terms of the penalty parameter C, it was defined as 10 since this value provided better performance (gains of 2/3% in every metric analysed when comparing to the standard of the library we used, which is a C parameter value of 1) while still having low variation in the metrics calculated in different executions of the model. This analysis of variation, like the one made for HMM, was done by assessing multiple executions of our model, and guaranteeing that the standard deviation of the results in every metric was not higher than 2%. For the gamma parameter, we chose not to tune it, given the already mentioned propensity for overfitting, and used the default value of the library we used, whose formula is of 1 divided by the number of features in the model. In our case, this value was 1/12, which, if we round up to thousandths, is 0.083.

With all of the information above, we contextualize our Results by saying that, despite these being reasonably positive, they should be taken with a grain of salt since the process that occurred in order to obtain the model mentioned included losing data regarding the last state given (which given that an individual has at most 4 states, is significant) and most of the thought process that lead to the definition of our observed states included an extensive amount of decisions and assumptions made in order to ensure the model is actually possible to be built and is logical according to our data as well as the question that we wished to answer with it.

In terms of the probabilities of *change* or *no change* given each variable, it is important to also put the values given in context. As seen by the values in Table 12, the probabilities of both *change* and *no change* when there was a *change of individual B* are of around 30 %. This means that overall, the number of inputs of our model that are estimated by the HMM to have *change of individual B* with a positive value are of about 30 %. This can be explained due to two important factors. The first is that this change variable consists of the same logic for our hidden state, *change*, but multiplied by the number of connections each individual has. For example, let's assume individual 23 has 3 answers and has relationships to 3 individuals while individual 42 has 3 survey answers as well and has relationships to 8 other individual. For *change*, the values for individual 23 have the same impact in the overall distribution of values of this variable as individual 42. For *change of individual B* however, the values of individual 42 have significantly more impact than the ones of individual 23 since individual 42 has more than double the amount of connections. Despite this, since as explained previously, the network we are using is highly connected, causing these differences between the actual probability of *change* and the estimated probability of *change of individual B* to not be significantly different since our hidden state that consists of *change* has similar a positive value rate. The second factor is that, before these calculations, while

defining our hidden state, we had already defined what consisted of *change* with one of the main goals being a balanced distribution. Therefore, before the probabilities are generated by our model, these results for the variable of *change of individual B* were already limited by the assumptions we had created while defining our model.

With both of these factors in mind, we reach the conclusion that there is strong evidence that the results in terms of influence of *change of individual B* in *change* calculated in Table 12 are heavily limited by the design of the network, as well as the definition of our model. As such, and considering that the difference between both probabilities of *change* or *no change* when considering individual B changed is relatively small (around 5%), the result given, while indicating slight influence, cannot lead to a defining conclusion on whether the individuals of the network of a specific individual truly influence that individual over time.

10. CONCLUSION

This thesis' main focus was to better understand how music preferences evolve over time and, when considering what it represents, music genre preference consists of a concept that is hard to quantify. Not only is it difficult for a person to allocate specific bands to a particular genre, a mere song can be interpreted as being from two different genres depending on the person that is listening to it. The same can be said for social relationships. Two people in a relationship can have a very different idea of what this connection consists of, and there is very fine line between a connection that consists of two people that occasionally socialize and actual friendship. As such, the clear difficulty in the definition of these concepts makes it ever so harder to define any relationship between them.

During our work, we had defined three main objectives. The first consisted in an analysis of a social network inherent to our data. To do this, we first executed a thorough analysis of not only the data we possessed regarding social relationships, but also of our musical preferences data, since we planned to use both in conjunction later on. During this analysis we found various interesting characteristics in each, but with one particular difference between them. While the social relationships dataset was balanced in terms of the timing of each surveys' response, the musical preferences dataset had multiple outliers in specific months.

This analysis proved crucial in our network definition, since this timing difference ultimately lead to us choosing to define our network as static, meaning that it would not vary over time. We then proceeded to use various SNA concepts, such as centrality metrics and possible sub-structures within a network that, combined with the representation we obtained, showcased just how connected this network of college students truly is.

The second goal was understanding how musical preferences evolve over time within the context of our network, with the goal of assessing if there are tendencies towards homogenization of music genre preferences over time. To do this, we merged our two different datasets and compared the percentages of differences of genre interest between individuals that are connected within the network over time. Having done this through two different criteria, one stricter and another more relaxed, we found that there are in fact tendencies of preferences between individuals that are connected being more similar over time. An additional analysis using the same type of data was also done, comparing differences between individuals with different levels of connection, yet we concluded that individual's with stronger connections do not seem to possess similar musical genre interests.

Our final goal consisted in going one step further and attempting to predict whether an individual changed preferences over time, when taking into account both the preferences of the individual himself as well as its connections' preferences within the network. To do this, we applied a first-order HMM using number of responses as our discrete points in time, whether an individual changed their binary musical preferences in at least 4 genres as our

states, and a combination of both an individual's preferences and whether one of his connections changed preferences at a particular point in time as our observations.

We compared this model with a SVM using the variables used in our combinations as its input and the *change* variable used in for states in HMM as our dependent variable. After comparing both of the results obtained using some of the more well-known classification metrics, both overall and using only the most common predictions for a particular individual at a particular point in time, we found that both models showed an interesting prediction capacity, with all previously referred metrics ranging between 0.65 and 0.9. We also analysed the probabilities present in the emission matrix to showcase how continuous the patterns of *change* and *no change* seem to be over time, as well as the probabilities of *change* when each of the binary variables are positive, obtained through our transition matrix and demonstrating the importance of each variable towards *change* in our model.

Lastly, we contextualize these results by explaining the entire thought process that went into the definition of each model, including the various experimentations done, ultimately leading to the conclusion that the results obtained both in terms of model strength and influence of variables aren't particularly conclusive, mostly due to the lack of data necessary to provide the most accurate results.

11. FUTURE WORK

As explained in the Experimentation and Discussion part of our work, despite the relatively positive results in terms of predictive ability of our model, due to the multiple difficulties found, whose root cause can be attributed to lack of data, these results should be taken with a grain of salt. As such, in this final chapter, we propose multiple paths that could be taken to further the work done within this thesis, including different types of solutions towards having more data, as well as strategies that could be used in order to take advantage of this potential additional data.

In terms of ways that we could possess more data, here are some possible solutions:

- Increase the timeframe of the study, including additional times that the surveys are taken by each individual. This change would lead to additional discrete points in time in our HMM, which could lead to better performance of this model.

- Having a larger number of individuals that respond in each survey. This would allow an increased number of rows in our dataset in exponential fashion, particularly if the high connectivity between individuals still remained, which could also lead better performance of our HMM model. This could also be particularly useful the additional individuals that responded the surveys were not college students, in order to help diversify the population of the network.

- Instead of just having survey data regarding specific genre preferences, possessing data regarding time spent by each individual listening to each type of music on a monthly/daily basis could also provide further insights in terms of different levels of genre preference in each individual within the network.

- Additional information regarding music artists, albums and particular songs listened to by each individual on a daily basis would perhaps provide more conclusions regarding music related behavior within the existing members of the network, as well as help shape the existing social network by taking into account possible connections if individuals are listening to the same songs at the same time.

- In terms of social network data, using existing location, proximity, and call logs, obtained through scanning nearby Wi-Fi access points and Bluetooth devices data to assess if, at a certain point in time, different individuals were spending time together could lead to a more accurate definition of the existing network, as well as a better understanding of how the network is evolving on a daily basis.

- Social media data regarding social interactions, such as number of tweets where each individual mentioned by another or number of Facebook / Instagram mentions could also provide additional insight regarding how the network is defined, as well as its evolution over time.

- Additional qualitative data in terms of feedback of each individual regarding the surveys that each took part could provide a better understanding of possible limitations of our data, which could lead to an overall better assessment of the evolution of musical preferences over time within the social network.

When considering possible strategies that could take the work started with this thesis to the level, the following applies:

- More data regarding the existing network, whether it is information regarding individuals' location or social media data, could allow for an assessment of the evolution of the existing network over time, which could provide better results of our model, as well as other interesting conclusions.

  Assessing the social network as directed or weighted (by considering, for example, number of Facebook mentions of an individual by another as the edge weight of the connection between those two individuals) could also be possible with additional data, without any significant loss of records within our model. This could also provide more insights and/or better results in our HMM.

- Finally, despite believing that the HMM is a type of model that suits our existing problem, different experiments using other predictive models that, for example, do not require a single variable that consists of combinations of others (such as the observable states in the HMM), could potentially provide not only better results, but additional information regarding the behavior of change of musical preferences within the network.

12. BIBLIOGRAPHY

Arif, T. (2015). The Mathematics of Social Network Analysis: Metrics for Academic Social Networks. *International Journal of Computer Applications Technology and Research*, *4*(12), 889–893. https://doi.org/10.7753/ijcatr0412.1003

Baum, L. E., & Petrie, T. (1966). Statistical inference for probabilistic funtions of finite state Markov chains. *The Annals of Mathematical Statistics*, *37*, 1554–1563.

Ben-Hur, A., & Weston, J. (2010). A user's guide to support vector machines. *Methods in Molecular Biology (Clifton, N.J.)*, *609*, 223–239. https://doi.org/10.1007/978-1-60327-241-4_13

Bicego, M., Grosso, E., & Otranto, E. (2008). A hidden Markov model approach to classify and predict the sign of financial local trends. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *5342 LNCS*, 852–861. https://doi.org/10.1007/978-3-540-89689-0_89

Butts, C. T. (2008). Social network analysis with sna. *Journal of Statistical Software*, *24*(6).

Christenson, P. G., & Peterson, J. B. (1988). Genre and Gender in the Structure of Music Preferences. *Communication Research*, *15*(3), 282-301. https://doi.org/10.1177/009365088015003004

Da Silva, A. P. A. (2014). Predicting market direction with hidden Markov models, (January).

Daniel, J., & Martin, J. H. (2018). Hidden Markov Models chapter A. *Speech and Language Processing*, 17.

Denora, T. (1999). Music as a technology of the self. *Poetics*, *27*(1), 31–56.

Dong, W., Lepri, B., & Pentland, A. (Sandy). (2011). Modeling the co-evolution of behaviors and social relationships using mobile phone data. *Proceedings of the 10th International Conference on Mobile and Ubiquitous Multimedia - MUM '11*, 134–143. https://doi.org/10.1145/2107596.2107613

Dong, W., Pentland, A., & Heller, K. A. (2012). Graph-Coupled HMMs for Modeling the Spread of Infection.

Easley, B. D. (2010). Networks, Crowds, and Markets: Reasoning about a Highly Connected World, *1*, 23–46.

Ferreira, A. (2013). Studying the impact of co-authorship with external researchers: the case of a research community in Portugal using Social Network Analysis.

Fletcher, A., Bonell, C., & Sorhaindo, A. (2011). You are what your friends eat: Systematic review of social network analyses of young people's eating behaviours and bodyweight. *Journal of Epidemiology and Community Health*, *65*(6), 548–555. https://doi.org/10.1136/jech.2010.113936

Frank, K. A. (1996). Mapping interactions within and between cohesive subgroups. *Social Networks*, *18*(2), 93–119. https://doi.org/10.1016/0378-8733(95)00257-X

Freeman, L. (2004), The Development of Social Network Analysis

Gingrich, P. (2004). Chapter 11 Association Between Variables. *Introductory Statistics for the Social Sciences*

Hidden Markov Models Simplified - Sanjay Dorairaj - Medium. (n.d.). Retrieved September 24, 2019, from https://medium.com/@postsanjay/hidden-Markov-models-simplified-c3f58728caab

Kupiec, J. (1992). Robust part-of-speech tagging using a hidden Markov model. *Computer Speech and Language*, *6*(3), 225–242. https://doi.org/10.1016/0885-2308(92)90019-Z

Lee, Yuh-Jye & Yeh, Yi-Ren & Pao, Hsing-Kuo. (2012). Introduction to Support Vector Machines and Their Applications in Bankruptcy Prognosis. Handbook of Computational Finance.

Hossin, M. & Sulaiman M. N. (2015). A Review on Evaluation Metrics for Data Classification Evaluations. *International Journal of Data Mining & Knowledge Management Process*, *5*(2), 01-11. https://doi.org/10.5121/ijdkp.2015.5201

North, A. C., & Hargreaves, D. J. (2012). The functions of music in everyday life: Redefining the social in music psychology. *Psychology of Music*, *27*(1), 71–83.

Preoţiuc-Pietro, D., & Cohn, T. (2013). Mining user behaviours: a study of check-in patterns in location based social networks. *A Study of Check-in Patterns in Location Based Social Networks*, 306–315. https://doi.org/10.1145/2464464.2464479

Rabiner, L. R. (1989). A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, *77*(2).

Rentfrow, P. J., & Gosling, S. D. (2003). The Do Re Mi's of Everyday Life: The Structure and Personality Correlates of Music Preferences. *Journal of Personality and Social Psychology*, *84*(6), 1236–1256. https://doi.org/10.1037/0022-3514.84.6.1236

Rentfrow, P. J., & Gosling, S. D. (2006). Message in a ballad: The role of music preferences in interpersonal perception. *Psychological Science*, *17*(3), 236–242. https://doi.org/10.1111/j.1467-9280.2006.01691.x

Rentfrow, P. J., & Gosling, S. D. (2007). The content and validity of music-genre stereotypes among college students. *Psychology of Music*, *35*(2), 306–326. https://doi.org/10.1177/0305735607070382

Rydén, T., Teräsvirta, T., & Åsbrink, S. (1998). Stylized facts of daily return series and the hidden Markov model. *Journal of Applied Econometrics*. https://doi.org/10.1002/(SICI)1099-1255(199805/06)13:3<217::AID-JAE476>3.0.CO;2-V

Saarikallio, S., & Erkkilä, J. (2007). The role of music in adolescents' mood regulation. In *Psychology of music* (Vol. 35, pp. 88–109). Sage Publications Sage CA: Thousand Oaks, CA.

Stamp, M. (2018). A Revealing Introduction to Hidden Markov Models. *Introduction to Machine Learning with Applications in Information Security*, 7–35. https://doi.org/10.1201/9781315213262-2

Suykens, Johan & Vandewalle, Joos. (1999). Least Squares Support Vector Machine Classifiers. Neural Processing Letters.

Tsvetovat, M., & Kouznetsov, A. (2011). *Social Network analysis for startups*.

Viterbi, A. J. (1967). Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm. *IEEE Transactions on Information Theory*, *13*(2), 260–269. https://doi.org/10.1109/TIT.1967.1054010

Wey, T., Blumstein, D. T., Shen, W., & Jordán, F. (2008). Social network analysis of animal behaviour: a promising tool for the study of sociality. *Animal Behaviour*, *75*(2), 333–344. https://doi.org/10.1016/j.anbehav.2007.06.020

Yamron, J. P., Carp, I., Gillick, L., Lowe, S., & Van Mulbregt, P. (1998). A hidden Markov model approach to text segmentation and event tracking. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, *1*(January 1992), 333–336. https://doi.org/10.1109/ICASSP.1998.674435